Title

Estimating the Reliability and Stability of Cognitive Processes Contributing to Responses on the Implicit Association Test.

Authors

Elder, Jacob
Wilson, Liz
Calanchini, Jimmy

Publication Date

2023-05-19

DOI

10.1177/01461672231171256

Peer reviewed

NOTE: This version of the manuscript has been accepted for publication in *Personality and Social Psychology Bulletin.*

Estimating the reliability and stability of cognitive processes contributing to responses on the Implicit Association Test

Jacob Elder[1], Liz Wilson[1], Jimmy Calanchini

University of California, Riverside

[1]The first two authors contributed equally to the preparation of this manuscript.

Address for correspondence: Jimmy Calanchini, Department of Psychology, 900 University Ave., University of California, Riverside, CA, 92521 Email: jimmy.calanchini@ucr.edu

TOTAL WORD COUNT: 11,034 (excluding title page, tables, and figures)

Abstract

Implicit measures were initially assumed to assess stable individual differences, but other perspectives posit that they reflect context-dependent processes. This pre-registered research investigates whether the processes contributing to responses on the race Implicit Association Test are temporally stable and reliably measured using multinomial processing tree modeling. We applied two models – the Quad model and the Process Dissociation Procedure – to six datasets (N = 2,036), each collected over two occasions, examined the within-measurement reliability and between-measurement stability of model parameters, and meta-analyzed the results. Parameters reflecting accuracy-oriented processes demonstrate adequate stability and reliability, which suggests these processes are relatively stable within individuals. Parameters reflecting evaluative associations demonstrate poor stability but modest reliability, which suggests that associations are either context-dependent or stable but noisily measured. These findings suggest that processes contributing to racial bias on implicit measures differ in temporal stability, which has practical implications for predicting behavior using the Implicit Association Test.

ABSTRACT WORD COUNT: 150

KEYWORDS: intergroup bias, implicit association test, racism, formal modeling, measurement reliability

Implicit measures were initially assumed to assess stable individual differences that reflect durable associations stored in memory (Fazio et al., 1995; Greenwald et al., 1998; Wilson et al., 2000)[1]. However, responses on implicit measures often demonstrate low temporal stability despite adequate reliability within measurement occasion (Bar-Anan & Nosek, 2014; Cunningham et al., 2001; Gawronski et al., 2017; Lai & Wilson, 2021). This pattern of findings begs the question: To what extent are the underlying processes that contribute to responses on implicit measures stable within individuals? We rely on multinomial processing tree models to disentangle the joint contributions of multiple cognitive processes to responses on the race Implicit Association Test (IAT: Greenwald et al., 1998), and we examine within-measurement reliability and between-measurement stability of parameters assumed to correspond to those latent processes. In doing so, we provide insight into the extent to which the cognitive processes that contribute to responses on the IAT are stable within individuals.

A variety of social cognitive theories assume that implicit measures primarily assess mental associations[2] between target categories (e.g., "ingroup") and attributes (e.g., "good") that are stored in memory and persist over time (e.g., Fazio, 1990, 2007; Petty et al., 2007; Strack & Deutsch, 2004). Indeed, implicit measures were developed with operating conditions (e.g., short response windows) intended to facilitate the expression of associations by minimizing the expression of cognitive processes that may vary as a function of motivation, opportunity, or other contextual factors (Fazio et al., 1995; Greenwald et al., 1998). To the extent that implicit

[1] In this manuscript, we use the term 'implicit' to mean 'indirect'. Thus, an 'implicit measure' assesses mental contents indirectly, in contrast to other forms of measurement that assess mental contents through direct inquiry. Our use of this term contrasts with other definitions that refer to the qualitative nature of the construct (e.g., unconscious), but aligns with the perspective that implicit measures assess evaluations under suboptimal processing conditions (De Houwer & Boddez, 2022). Here, we specify indirect measurement as the defining procedural feature of the task that is the focus of the present research (i.e., the Implicit Association Test; Greenwald et al., 1998).

[2] In this manuscript, we use the term 'association' to refer to one of the mental constructs that influences responses on implicit measures. However, we adopt the term largely in recognition of its conventional use, but we make no strong assumptions about the representational nature – associative or otherwise – of the construct.

measures assess associations that are stable and enduring within persons, then responses on implicit measures should be expected to predict individual behaviors. Several meta-analyses have tested this assumption, and they estimated small-to-moderate relationships between the IAT and behavioral outcomes (Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). Like any other measurement tool, implicit measures must assess the intended construct reliably in order to effectively predict behaviors and other individual differences (Kanyongo et al., 2007; Loken & Gelman, 2017; Schmidt & Hunter, 1996). However, the IAT in particular has been criticized as a noisy measure (Blanton et al., 2009; Schimmack, 2021) due to its low retest stability across measurement occasions (Bar-Anan & Nosek, 2014; Cunningham et al., 2001; Gawronski et al., 2017; Lai & Wilson, 2021). To some degree, these criticisms depend on the assumption that the construct assessed by implicit measures is a stable individual difference – an assumption that is not universally accepted. For example, the Bias of Crowds (Payne et al., 2017) proposes that variance in responses on implicit measures is better explained by differences in situations and contexts than by differences between people. This contextual perspective dovetails with constructivist attitude theories (e.g., Conrey & Smith, 2007; Schwarz, 2007), which propose that responses on attitude measures (implicit or otherwise) do not reflect anything stable within individuals but, instead, reflect evaluations that are constructed on-the-spot based on information that is momentarily accessible – either in the mind or in the environment. To the extent that responses on implicit measures reflect constructs that are situationally dependent, then low retest stability across measurement occasions is unsurprising.

The unresolved debate over whether implicit measures assess something stable versus context-dependent largely depends on the assumptions that responses on implicit measures are relatively process-pure and primarily reflect the influence of mental associations. However,

research using multinomial processing tree models (MPTs: Riefer & Batchelder, 1988) indicates that multiple cognitive processes jointly contribute to responses on implicit measures (Calanchini, 2020; Hütter & Klauer, 2016). Responses on implicit measures are traditionally quantified using summary statistics (e.g., Greenwald et al., 2003), but summaries can only provide limited insight when multiple processes influence responses. Thus, the extent to which responses on implicit measures reflect some processes that are stable within individuals and others that are not remains an open question. Compared to traditional summary statistics, MPT models are well-positioned to provide relatively more theoretically precise and statistically rigorous insight into the cognitive processes that contribute to responses on implicit measures.

In the present research, we examine the extent to which the cognitive processes that contribute to responses on the race IAT reflect stable individual differences versus context-dependent processes. As an analog to how we conceptualize stability, people exhibit stability of personality traits across time (Caspi & Bem, 1990; Roberts, Wood, & Caspi, 2008; Caspi, Bem, & Elder, 1989), such that they often maintain some degree of relative rank-ordering of behavior, regardless of average changes in the personality trait or behavior over time (i.e., the average may change, but the ordering is consistent). In the same way, we assume that some cognitive processes that contribute to responses on the IAT may be more or less stable across time. To examine stability in these processes, we examined the consistency of MPT parameters estimated from IATs administered across two measurement occasions. In psychometrics, retest stability metrics are used to parse between-individual variability in responses that represents "true" scores from within-individual variability that represents measurement error. However, within-individual variability across measurement occasions does not necessarily reflect measurement error if the underlying construct is context-dependent (Röseler et al., 2020; Steyer et al., 1992; Zuckerman,

1983). To determine whether low consistency across measurement occasion reflects instability in the underlying process versus measurement error, we must first establish whether the measure is reliable within measurement occasions. In order to help shed light on the extent to which IAT responses reflect within-individual variability versus error across measurement occasions, we also estimated the reliability of MPT parameters within each measurement occasion using parameter recovery. In summary, we establish within measurement reliability using parameter recovery measurement, and we establish between measurement stability using retest consistency measurement.

In this manuscript, we adopt the following terminology to distinguish between the tests we perform and the inferences we draw from those tests about the underlying constructs. We assess *consistency* (rather than absolute agreement, which accounts for systematic differences across timepoints) in parameter estimates across measurement occasions using intra-class correlations (ICCs) to draw inferences about the *stability* with which the constructs reflected in the model parameters can be measured. We also assess the *recoverability* of parameters within measurement occasions using parameter recovery to draw inferences about how *reliably* the constructs are measured by the model parameters.

Taken together, the analytic approach we adopt in the present research consists of two primary sets of analyses. To assess within-measurement reliability, we simulated data to determine the extent to which each MPT parameter can be reliably recovered. To assess between-measurement stability, we estimated the consistency of MPT parameters across occasions using intraclass correlations (ICCs). We will interpret parameters that demonstrate acceptable consistency and recoverability to reflect reliably measured and stable cognitive processes (i.e., trait individual differences), and interpret parameters that demonstrate poor

consistency but acceptable recoverability to reflect reliably measured but unstable processes (i.e., time-dependent states). Parameters that demonstrate poor consistency and recoverability are unstable and unreliably measured, and therefore unlikely to be valid measures for individual-level inference. To increase the validity of our findings, we repeat this procedure across six independent datasets and meta-analyze the results. Moreover, to further maximize the validity of our findings, we apply two different MPT models (e.g., the Quad model and the Process Dissociation Procedure) to each dataset and look for consistent patterns of results in conceptually analogous MPT parameters.

Method

Study Selection

We relied on one dataset collected in our lab, and five other datasets from other sources. All six datasets consisted of data from the race IAT administered to participants on two measurement occasions. Sample sizes range from n = 32-1,240 participants, and intervals between measurement occasions range from a few minutes to two years (Table 1). Most of the data were collected online, so we do not have information about the physical locations in which the two measures were completed. Consequently, our analyses are well-positioned to provide insight into the extent to which processes reflect stable individual differences, but they provide insight only into the temporal dimension of context-dependence.

The present research was approved by the Institutional Review Board of the University of California Riverside #HS 20-278. Because this research relies on existing datasets, we did not conduct power analyses or determine sample sizes based on the present research questions. Similarly, these datasets may have included other manipulations or measures that are not relevant to our research questions, so we do not report or analyze those here. We describe our exclusion

criteria below. Unless otherwise noted, all hypotheses and analyses were pre-registered. Pre-registrations, code, and data from Wilson & Calanchini (2022), Forscher et al., (2017), and Project Implicit (2020) are available at https://osf.io/qgvz3/?view_only=56aaa617de3545c297a5a1ce35b79ed2. Data from Lai et al. (2016) are available at https://osf.io/dbtns/. Data from Gawronski et al., (2017) are available at https://osf.io/792qj/.

Table 1. *Description of datasets*

Source	N	Approximate Time Interval
Project Implicit, 2020	1240	One browser session[3]
Wilson & Calanchini, 2022	105	24 - 48 hours
Lai et al., 2016 (Study 1)	80	1 - 4 days
Lai et al., 2016 (Study 2)	463	1 - 4 days
Gawronski et al., 2017	116	1 month
Forscher et al., 2017	32	2 years

The Race Implicit Association Test

All studies relied on the race IAT, which consists of stimuli reflecting two target categories (Black, White) and two attribute categories (good, bad[4]). The IAT proceeds in seven blocks, with the first block consisting of 20 practice trials in which participants categorize White stimuli and Black stimuli using two computer keys. The second block consists of 20 practice trials in which participants categorize good words and bad words. In the third and fourth critical

[3] The Project Implicit (2020) dataset consists of participants who completed two race IATs within the same browser session, but it does not record the time interval between IATs.

[4] IAT attribute category labels were slightly different across datasets. The Project Implicit, Wilson and Calanchini, and Lai datasets used "good" and "bad" as category labels. Gawrsonki used "positive" and "negative" as category labels. Forscher used "pleasant" and "unpleasant" as category labels.

blocks, the stimuli and response keys are combined, such that participants complete a total of 60 trials in which they respond to White and good stimuli with one response key, and to Black and bad stimuli with the other response key. The fifth block consists only of Black and White stimuli, and participants complete 20 practice trials with the response mapping reversed relative to the mapping in the previous blocks. In the sixth and seventh critical blocks, participants completed a total of 60 trials in which they respond to White and bad stimuli with one response key, and to Black and good stimuli with the other response key. Both of the Lai et al. (2016) datasets slightly deviated from this task structure. In these studies, participants completed an abbreviated version of the IAT with five blocks instead of seven blocks. Rather than four critical blocks, the abbreviated version consisted of two critical blocks of 32 trials each.

Data Pre-Processing

We excluded participants who did not complete all IAT critical trials at both measurement occasions from analysis. Additionally, we excluded participants who demonstrated IAT errors rates exceeding 50%, which corresponds to random responding.

From the Gawronski et al. (2017) dataset, 4 participants' responses were unable to be matched between measurement occasions due to an error in subject identifiers. From the Wilson and Calanchini (2022) dataset, 27 participants were excluded from analysis due to not fully completing the IAT at both measurement occasions. We examined a subset of respondents in the Project Implicit (2020) dataset and excluded participants who did not fully complete two separate IATs within the same browser session. After this exclusion, 12 additional participants in the Project Implicit (2020) dataset were removed for error rates exceeding 50%. After applying these exclusion criteria, we were left with the final sample sizes reported in Table 1.

Multinomial Processing Tree Models

Responses on the IAT are traditionally quantified according to the D-scoring algorithm (Greenwald et al., 2003), which is a summary statistic that reflects the standardized difference between participants' response latencies to one block of trials (e.g., when White stimuli share a response key with good attributes) versus another block of trials (e.g., when Black stimuli share a response key with good attributes). In the context of the race IAT, D-scores are interpreted such that values greater than zero are assumed to reflect relatively more positive evaluations of White people, and values less than zero are assumed to reflect relatively more positive evaluations of Black people. However, operationalizing responses on the IAT in terms of a relative summary statistic is theoretically imprecise: for example, differences in D-scores between experimental conditions may indicate that responses on the IAT are sensitive to manipulation but does not provide insight into which cognitive process or processes were influenced by the manipulation. In contrast, MPT modeling (Riefer & Batchelder, 1988) provides greater theoretical precision than do D-scores by quantifying the joint contributions of multiple cognitive processes to responses.

MPT models belong to a class of formal mathematical models that link latent processes to observable responses on tasks like the IAT (Batchelder & Riefer, 1999). MPT models 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 contribute to responses in the paradigm (Hütter & Klauer, 2016). In creating MPT models, researchers must make theoretically grounded decisions about the specific manner in which multiple cognitive processes produce responses in each task condition. In this way, MPT models are mathematical instantiations of psychological theory packaged in a well-defined form.

An MPT model consists of parameters that correspond to latent cognitive processes, and the proposed interplay of these processes 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 the preceding process. 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.

In the present research, we relied on two well-validated MPT models that have frequently been applied to the IAT: the quadruple process model (Quad model; Conrey et al., 2005) and the process dissociation procedure (PDP; Payne, 2001). The two models share a common dual-process perspective on implicit social cognition but differ in the number of processes proposed to influence responses, as well as in assumptions about the qualitative nature of those processes. In the present research, we applied both MPT models to the same IAT data, which not only provides a conceptual replication of our tests across models, but also prevents us from making inferences based on a single operationalization of the cognitive processes underlying IAT performance. We can draw relatively stronger conclusions from our data if the same pattern of results emerges from both models.

The Quad Model

The Quad model is depicted in Figure 1, and posits that observable responses in the IAT are produced by the joint influence of qualitatively distinct cognitive processes reflected in four model parameters (Conrey et al., 2005).

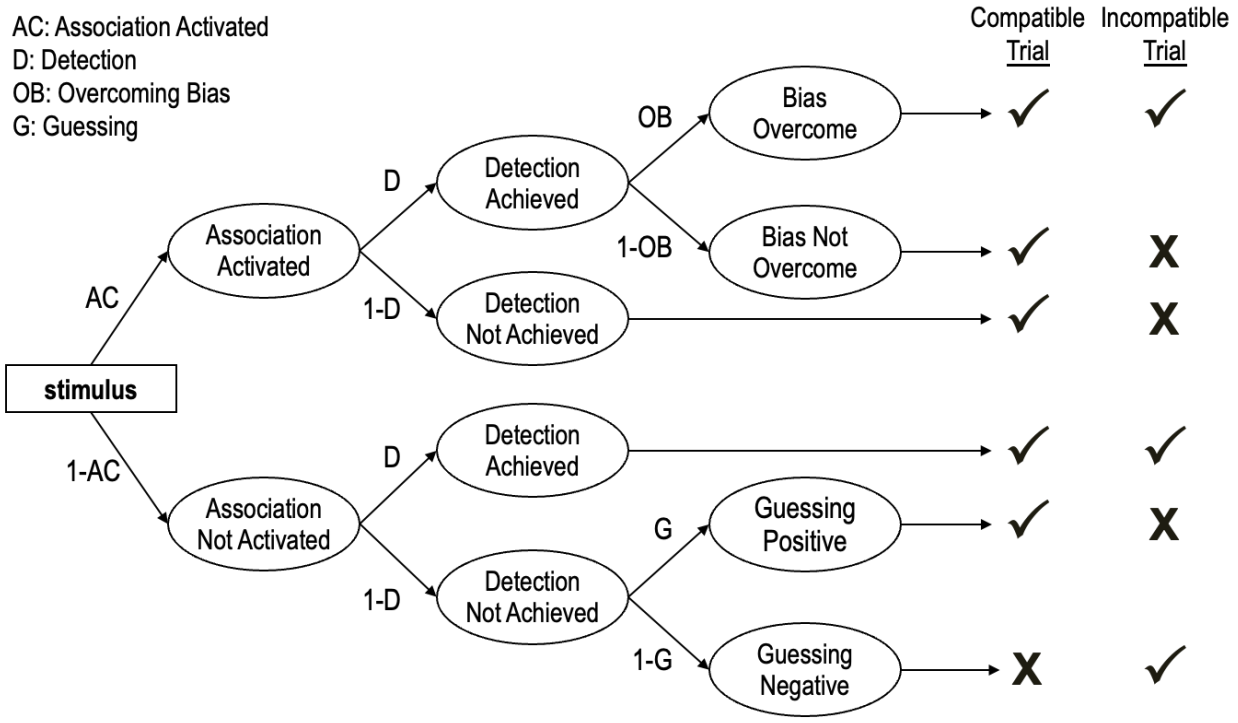

Figure 1. A portion of the Quad mode. The table on the right illustrates correct (✓) and incorrect (**X**) responses across different trial types.

The activation of Associations parameter refers to the degree to which mental associations are activated when responding to a stimulus. All else being equal, the stronger the association between the target (e.g., White) and the attribute (e.g., good), the more likely the association is to be activated and drive responses in an association-consistent direction. We estimated two different Associations parameters: one reflects an association between White and good, and the other reflects an association between Black and bad. The Detection of correct responses parameter is conceptualized as an accuracy-oriented process, and it reflects the likelihood that the participant can discern the correct response. Sometimes activated associations conflict with the detected correct response. For example, on trials in which White faces appear and the categories "White" and "bad" share a response key, to the extent that a participant associates "White" with "good" then activated associations would conflict with the detected

correct response. The Quad model proposes an Overcoming Bias parameter to resolve such a conflict between Associations and Detection. The Overcoming Bias parameter refers to an inhibitory process that prevents activated associations from influencing behaviors when they conflict with detected correct responses. Finally, the Guessing parameter does not represent a specific process, per se, but instead reflects the tendency to respond with "good" versus "bad" in the absence of influence from the Associations, Detection, and Overcoming Bias parameters.

The Process Dissociation Procedure

The PDP is depicted in Figure 2 and posits that observable responses on the IAT are produced by the joint influence of qualitatively distinct cognitive processes reflected in two model parameters.[5]

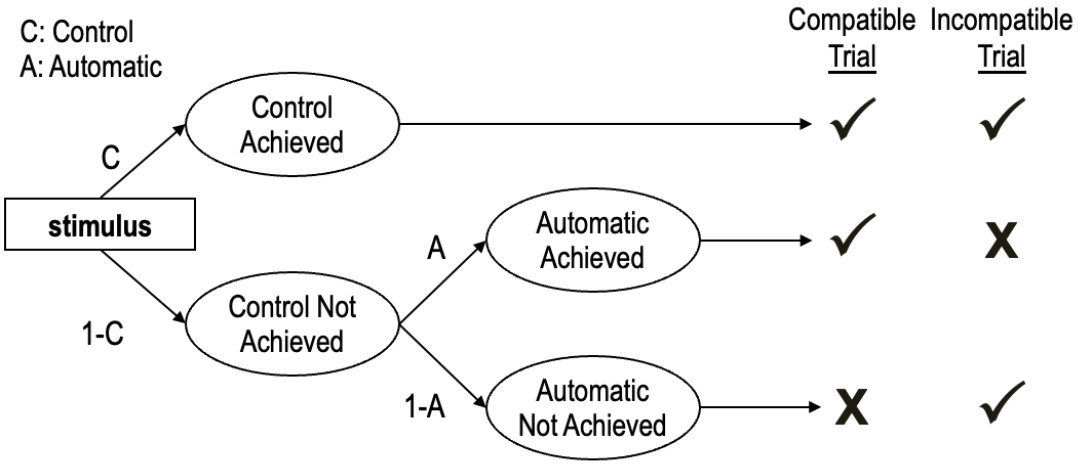

Figure 2. The Process Dissociation Procedure. The table on the right depicts correct (✓) and incorrect (**X**) responses across different trial types.

[5] The parameter names 'Automatic' and 'Control' correspond to the dual-process perspective of cognition (e.g., Metcalfe & Mischel, 1999; Schneider & Shiffrin, 1977) in which the field of implicit social cognition is deeply rooted. Unfortunately, these terms have become linked over the years with assumptions about the awareness, intentionality, controllability, efficiency, and speed of processes described as automatic or controlled -- assumptions that have not always been backed by empirical support (Melnikoff & Bargh, 2018; Moors & De Houwer, 2006). Consequently, we retain these labels to describe the two parameters of the PDP for linguistic convenience, but expressly make no assumptions about the qualitative nature (e.g., unawareness) of the cognitive process or processes that either parameter reflects beyond what the analyses reported in the present research can support.

The Automatic parameter refers to the degree to which associations between targets and attributes drive responses. Similar to how the Associations parameter of the Quad model is assumed to operate, the stronger the association between a target (e.g., White) and an attribute (e.g., good), the more likely the association is to be activated and drive responses in an association-consistent direction. We estimated two different Automatic parameters, one for White stimuli and another for Black stimuli. Importantly, though the Automatic parameters of the PDP and the Associations parameters of the Quad model are conceptually analogous, they are operationalized differently in each model. The Associations parameters of the Quad model assume a direction of compatibility, such that they are traditionally specified to reflect associations between White people and good attributes, and between Black people and bad attributes; consequently, larger parameter values reflect stronger links between White-good and Black-bad. In contrast, the Automatic parameters of the PDP do not assume a direction of compatibility: parameter values greater than 0.5 reflect positive evaluations and values less than 0.5 reflect negative evaluations of each target group.

The Control parameter reflects correct processing of stimuli and how well participants can distinguish between target concepts and attributes. The Control parameter in the PDP and the Detection parameter in the Quad model are conceptually analogous, in that both refer to accuracy-oriented cognitive processes. However, the two parameters differ in their specifications. In the PDP, the success of the Control parameter will always lead to a correct response, and the Automatic parameter can only influence responses in the absence of influence from the Control parameter. In contrast, in the Quad model, the success of the Detection parameter depends on the success of the Overcoming Bias parameter when the Associations parameter would produce a response that conflicts with Detection. We estimated four different

Control parameters, one each for White, Black, good, and bad stimuli.

Because the PDP does not include a catch-all parameter like Guessing in the Quad model, the Automatic parameters of the PDP must be interpreted differently from the Associations parameters of the Quad model. Specifically, the influence of any cognitive processes that are not accounted for by the Control parameter is necessarily reflected in the Automatic parameter. The PDP assumes that the Control parameter will always drive responses, even if the Automatic parameter is also activated. In contrast, the Quad model assumes that either the Detection parameter or the Associations parameter can drive responses if both are activated, and the success or failure of the Overcoming Bias parameter determines whether Detection or Associations drive responses, respectively. Consequently, the Automatic parameter of the PDP can be interpreted to reflect a combination of the Associations, Overcoming Bias, and Guessing parameters of the Quad model.

Interpretation of Processes

Because the Associations (in the Quad model) and Automatic (in the PDP) parameters are conceptualized to reflect associations between target groups (e.g., White) and evaluations (e.g., good), they would seem to most closely correspond to the construct that implicit measures are traditionally assumed to assess. If these parameters demonstrate fair consistency between measurement occasions and acceptable recoverability within measurement occasions, we will interpret them to reflect a reliable measure of stable individual differences. However, if they demonstrate poor consistency but acceptable recoverability, we will interpret them to reflect a reliable measure of a time-dependent process.

The other parameters reflected in the Quad model and PDP do not correspond as closely as do the Associations and Automatic parameters to the constructs that implicit measures are

traditionally assumed to assess. Nevertheless, these other parameters are estimated from IAT responses, so investigating the extent to which they are reliable and stable across measurement occasions may still be informative. The Detection and Control parameters are conceptualized to reflect accuracy-oriented processes, and the Overcoming Bias parameter is conceptualized to reflect an inhibitory process. Guessing does not reflect a specific process, but instead reflects any processes that influence IAT responses that are not accounted for by the other model parameters, akin to residual error terms in structural equation modeling. We will interpret these parameters in the same way that we interpret the Associations and Automatic parameters: parameters that demonstrate fair consistency and acceptable recoverability reflect reliable measures of stable processes, and parameters that demonstrate poor consistency but acceptable recoverability reflect processes that are reliable measures of time-dependent processes. A parameter with poor recoverability is likely to also have poor consistency and/or contain substantial measurement error and, thus, should not be relied on for individual-level inference.

Multinomial Processing Tree Estimation

To quantify the influence of each process specified in each MPT model, we implemented an approach that relies on hierarchical Bayesian estimation (Klauer, 2010). This approach assumes that individual-level parameters are drawn from a multivariate normal population distribution, thereby regularizing and stabilizing individual-level estimates (Ahn et al., 2017). We fitted all hierarchical MPT models using default priors in *TreeBUGS* (Heck et al., 2018) in R Programming Environment v.4.1.2., which draws posterior samples of the parameters using Markov chain Monte Carlo methods. We ensured sufficient parameter convergence of all models using a criterion of a Gelman-Rubin R-hat < 1.05 and a visual inspection of Gelman-Rubin trace plots.

Pre-registered Analyses

Parameter Consistency Between Measurement Occasion

We assessed parameter consistency between measurement occasions using ICCs representing the ratio of intra- to inter-individual variance (Koo & Li, 2016; McGraw & Wong, 1996). We modeled two-way mixed effects ICCs, such that both participants and measurement occasions were treated as random effects sampled from a larger pool of people and time points. We focused on the relative ranking of participants across timepoints, rather than absolute agreement without error, and thus relied on consistency ICCs. Each ICC was conducted with only two timepoints, and we report ICCs with confidence intervals (CIs). We defined as ICC(3,1) according to Shrout & Fleiss (1979) convention[6], and estimated ICCs using the *irr* package 0.84.1 (Gamer et al., 2010).

Parameter Recovery Within Measurement Occasion

Parameter recovery is a method to investigate the extent to which a specific model configuration can reliably reproduce parameter estimates given a set of behavioral data (i.e., the model is identifiable). In doing so, parameter recovery provides us with insight into the extent to which a parameter reflects an estimate of the intended construct versus measurement error (Ballard et al., 2020; Shahar et al., 2019). The parameter recovery process consists of four steps. First, we estimate a set of model parameters from real participants' responses (i.e., "original" parameters). Second, we simulate behavioral data based on the original parameters. Third, we fit the model to the simulated data to produce a new set of parameter estimates (i.e., "simulated" parameters). Fourth, and finally, we compare the simulated parameters to the original parameters. If a model's parameters can be successfully recovered, there will be tight

[6] The McGraw & Wong (1996) convention would be ICC(C,1).

correspondence between original and simulated parameters – which, in turn, provides a "ground truth" to establish the reliability of parameter estimation. As an analogy, parameter recovery in this context can be considered akin to a psychometric investigation of internal consistency, which may be more familiar to many readers. Internal consistency reflects the extent to which items on an inventory are correlated with one another and, thus, quantifies whether the measurement of a construct can be trusted. Similarly, parameter recovery reflects the extent to which a set of parameter values can generate behavior that reproduces the same parameters and, thus, quantifies whether the measurement of a parameter can be trusted.

We simulated behavioral data using the *rmultinorm* function in R based on the original Quad model and PDP parameters estimated from each dataset. The behavioral data corresponded to two choice outcomes – correct, incorrect – for each response category (i.e., responses to White, Black, pleasant, and unpleasant stimuli in compatible and incompatible IAT blocks). Then, we applied the Quad model and PDP to the simulated data and estimated new parameters. Finally, we calculated Pearson correlations between the original and recovered estimate of each parameter for each model for each dataset and time point.

Meta-Analysis

To synthesize our findings across datasets, we performed random-effects meta-analyses using the *metafor* package 3.0-2 in R (Viechtbauer, 2010). For each parameter of each MPT model, we conducted one meta-analysis based on the consistency analyses, and another meta-analysis based on the recovery analyses. Whereas the consistency analyses necessarily reflected data from both measurement occasions (i.e., quantifying the extent to which responses at Time 1 correspond with responses at Time 2), the recovery analyses reflected data within each measurement occasion, thereby providing twice as many estimates in the recovery meta-analysis

as in the consistency meta-analysis. Consequently, we modeled measurement occasion as a random factor nested within study in a multilevel meta-analysis of the recovery estimates.

We estimated the standard error of all ICCs using the Fisher r-to-Z transformation for ICC values (Chen et al., 2018). For the final reported meta-analytic estimates, we converted the estimates and their 95% confidence intervals (CIs) via a Z-to-r transformation. As inference criteria, we compare CIs to determine if meta-analytic estimates are significantly different from one another. For example, if the CIs of two parameters' meta-analytic ICCs do not overlap (i.e., the ICC's lower bound for one parameter is greater than the upper bound for another parameter), we will conclude that the two parameters' ICCs are different from one another.

Results of Pre-registered Analyses

Parameter Consistency Between Measurement Occasions

In order to estimate whether parameters are consistent between measurement occasions, we calculated ICCs for Quad and PDP parameters and meta-analyzed the results. Between-measurement occasion consistency is often interpreted using different criteria than is within-measurement occasion reliability (Matheson, 2019), given that changes may reflect changes in true scores or measurement error. We interpret ICCs according to the criteria proposed by Cicchetti & Sparrow (1981): < .40 is poor; .40 to .60 is fair; .60 to .75 is good; > .75 is excellent.

The Quad Model

ICCs for each Quad parameter for each dataset are depicted in Figure 3. Meta-analytic results (Figure 4) indicate that Detection parameters were the most consistent across measurement occasions of all Quad parameters, and demonstrate fair consistency, ICC(3,1) = .515, 95% CI = [.436 - .587], p < 001. The other parameters demonstrated poor consistencies: White-good Associations ICC(3,1) = .318, 95% CI = [.170 - .452], p < .001; Black-bad

Associations ICC(3,1) = .104, 95% CI = [.061 - .147], p < .001; Overcoming Bias ICC(3,1) = .160, 95% CI = [.026 - .289], p = .020; Guessing ICC(3,1) = .012, 95% CI = [-.124 - 148], p = .863. The consistency of the Guessing parameter approaches 0 and its confidence interval includes negative values, indicating that its within-subject variance exceeds its between-subject variance, and suggesting that Guessing reflects more noise than signal.

Inspection of CIs indicates that Detection exhibits higher consistency than do Black-bad Associations, Overcoming Bias, and Guessing, but does not differ from the consistency of White-good Associations. White-good Associations demonstrate higher consistency than Guessing, but does not differ from the consistency of Black-bad Associations or Overcoming Bias. Black-bad Associations, Overcoming Bias, and Guessing do not differ in their consistency.

There was substantial heterogeneity in consistency between-studies for Detection ($Q(5) = 13.717$, $p = .018$, $I^2 = 67.626$), Guessing ($Q(5) = 50.089$, $p < .0001$, $I^2 = 82.406$), Overcoming Bias ($Q(5) = 18.887$, $p = .002$, $I^2 = 82.125$), and White-good Associations ($Q(5) = 32.507$, $p < .0001$, $I^2 = 86.163$), which suggests that these parameters varied in their ICCs across studies. However, there was minimal variation in Black-bad Associations ICCs across studies, $Q(5) = 6.819$, $p = .235$, $I^2 = 0$.

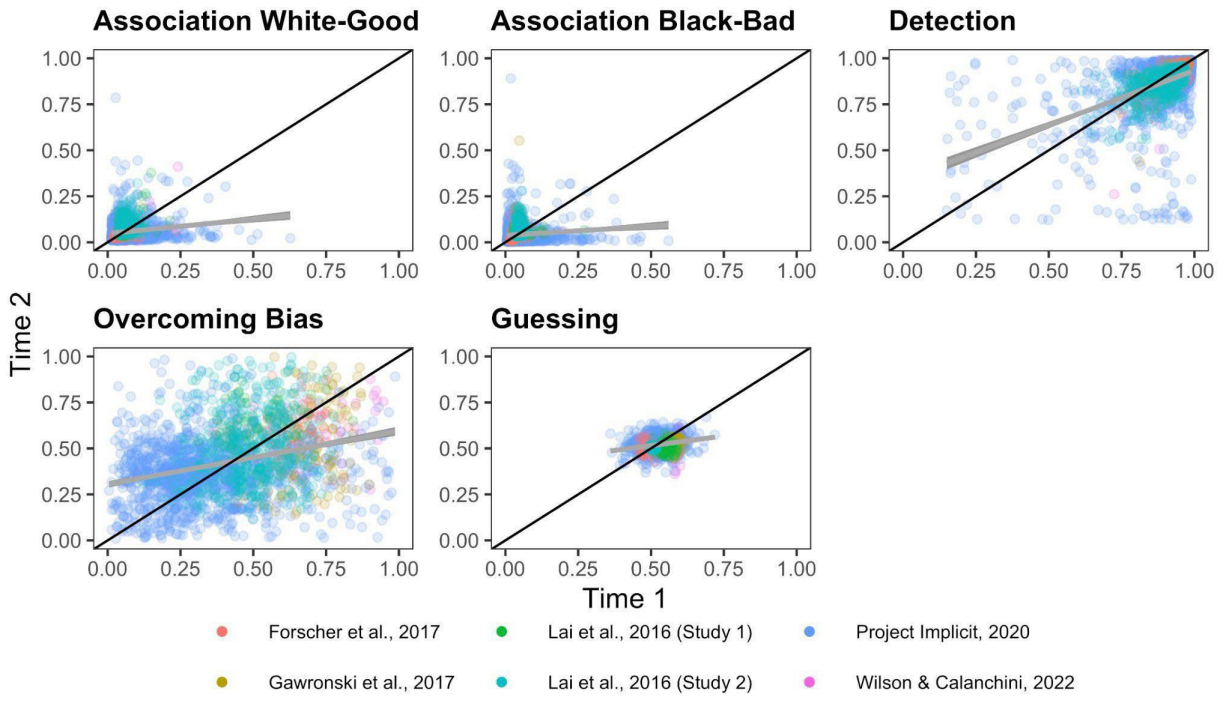

Figure 3. Scatterplots depicting the relationship between Time 1 and Time 2 Quad parameters. The black diagonal line depicts perfect consistency between Time 1 and Time 2 and the gray line depicts best-fit slopes through the observed data.

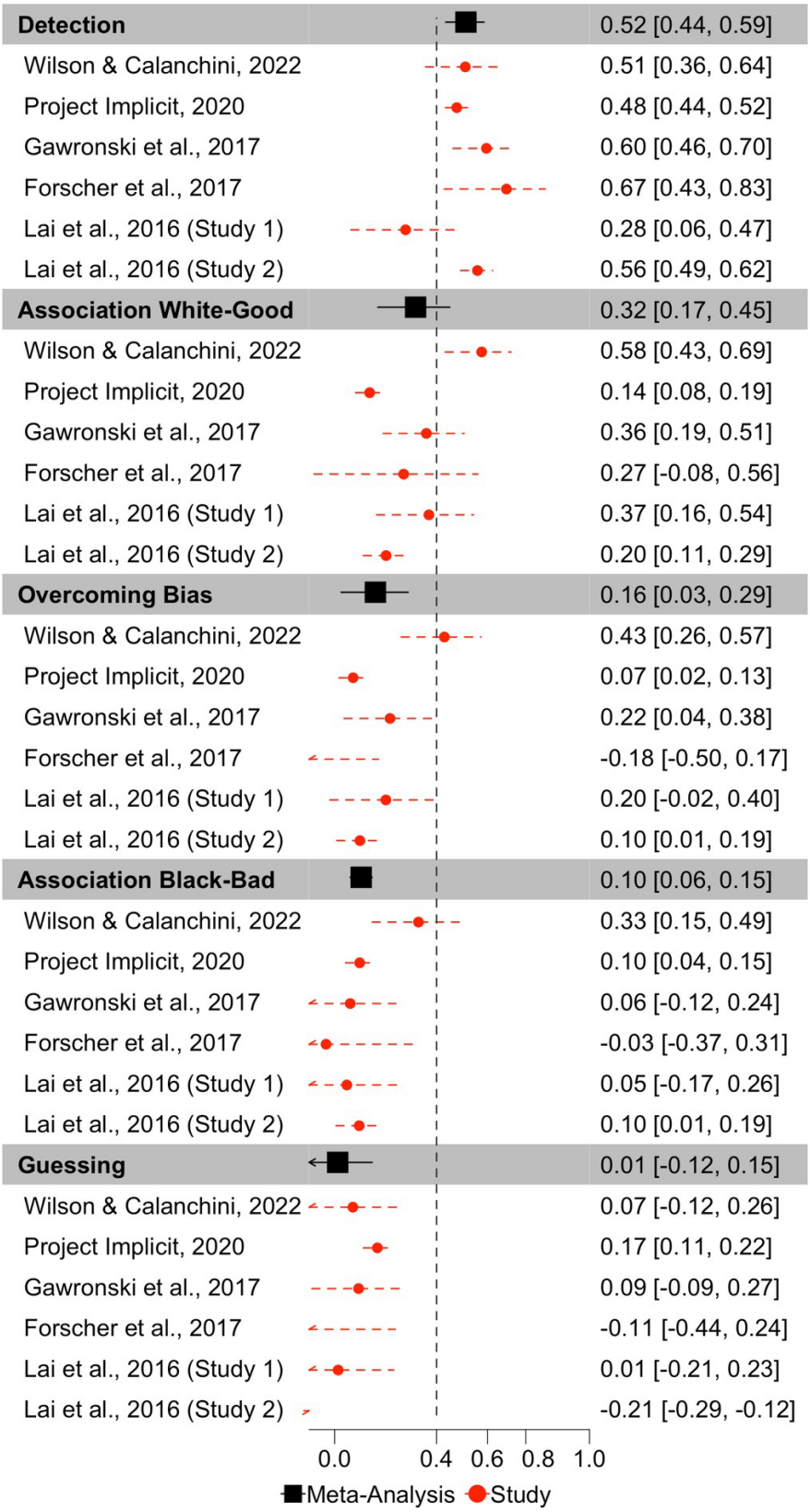
Detection
0.52 [0.44, 0.59]
Wilson & Calanchini, 2022
0.51 [0.36, 0.64]
Project Implicit, 2020
0.48 [0.44, 0.52]
Gawronski et al., 2017
0.60 [0.46, 0.70]
Forscher et al., 2017
0.67 [0.43, 0.83]
Lai et al., 2016 (Study 1)
0.28 [0.06, 0.47]
Lai et al., 2016 (Study 2)
0.56 [0.49, 0.62]
Association White-Good
0.32 [0.17, 0.45]
Wilson & Calanchini, 2022
0.58 [0.43, 0.69]
Project Implicit, 2020
0.14 [0.08, 0.19]
Gawronski et al., 2017
0.36 [0.19, 0.51]
Forscher et al., 2017
0.27 [-0.08, 0.56]
Lai et al., 2016 (Study 1)
0.37 [0.16, 0.54]
Lai et al., 2016 (Study 2)
0.20 [0.11, 0.29]
Overcoming Bias
0.16 [0.03, 0.29]
Wilson & Calanchini, 2022
0.43 [0.26, 0.57]
Project Implicit, 2020
0.07 [0.02, 0.13]
Gawronski et al., 2017
0.22 [0.04, 0.38]
Forscher et al., 2017
-0.18 [-0.50, 0.17]
Lai et al., 2016 (Study 1)
0.20 [-0.02, 0.40]
Lai et al., 2016 (Study 2)
0.10 [0.01, 0.19]
Association Black-Bad
0.10 [0.06, 0.15]
Wilson & Calanchini, 2022
0.33 [0.15, 0.49]
Project Implicit, 2020
0.10 [0.04, 0.15]
Gawronski et al., 2017
0.06 [-0.12, 0.24]
Forscher et al., 2017
-0.03 [-0.37, 0.31]
Lai et al., 2016 (Study 1)
0.05 [-0.17, 0.26]
Lai et al., 2016 (Study 2)
0.10 [0.01, 0.19]
Guessing
0.01 [-0.12, 0.15]
Wilson & Calanchini, 2022
0.07 [-0.12, 0.26]
Project Implicit, 2020
0.17 [0.11, 0.22]
Gawronski et al., 2017
0.09 [-0.09, 0.27]
Forscher et al., 2017
-0.11 [-0.44, 0.24]
Lai et al., 2016 (Study 1)
0.01 [-0.21, 0.23]
Lai et al., 2016 (Study 2)
-0.21 [-0.29, -0.12]
0.0
0.4
0.6
0.8
1.0
Meta-Analysis
Study

Figure 4. Forest plot depicting the meta-analytic and study specific consistencies for the Quad model, depicted as intraclass correlations. Point estimates and 95% confidence intervals are reported. Dashed vertical line reflects the threshold of acceptable consistency.

The Process Dissociation Procedure

ICCs for each PDP parameter for each dataset are depicted in Figure 5. Meta-analytic results (Figure 6) indicate that Control parameters were the most consistent across measurement occasions of all PDP parameters, exhibiting fair consistency: Control-good ICC(3,1) = .487, 95% CI = [.428 - .578], p < 001; Control-bad ICC(3,1) = .483, 95% CI = [.377 - .576], p < .001; Control-Black ICC(3,1) = .477, 95% CI = [.362 - .578], p < .001; Control-White ICC(3,1) = .440, 95% CI = [.283 - .574], p < .001. However, both Automatic parameters demonstrated poor consistency: Automatic-White ICC(3,1) = .233, 95% CI = [.082 - .374], p = .003; Automatic-Black ICC(3,1) = .232, 95% CI = [.069 - .382], p = .006.

Inspection of confidence intervals indicates that Control-good exhibits significantly higher consistency than do both of the Automatic parameters. Control-bad also exhibits significantly higher consistency than does Automatic-White, but does not differ in consistency from Automatic-Black. The consistency of the other two Control parameters does not differ from the consistency of either of the Automatic parameters. None of the Control parameters differ from one another in terms of consistency, nor do the Automatic parameters differ from one another in terms of consistency.

There was substantial heterogeneity in consistency between-studies for Automatic-Black ($Q(5) = 25.506$, $p = .0001$, $I^2 = 88.608$), Automatic-White ($Q(5) = 21.914$, $p = .0005$, $I^2 = 86.758$), Control-Black ($Q(5) = 15.738$, $p = .008$, $I^2 = 83.326$), Control-bad ($Q(5) = 15.580$, $p < .0001$, $I^2 = 80.419$), Control-White ($Q(5) = 26.757$, $p < .0001$, $I^2 = 90.539$). Control-good was the

only parameter which did not significantly vary in terms of consistency across studies, $Q(5) = 8.611$, $p = .126$, $I^2 = 42.710$.

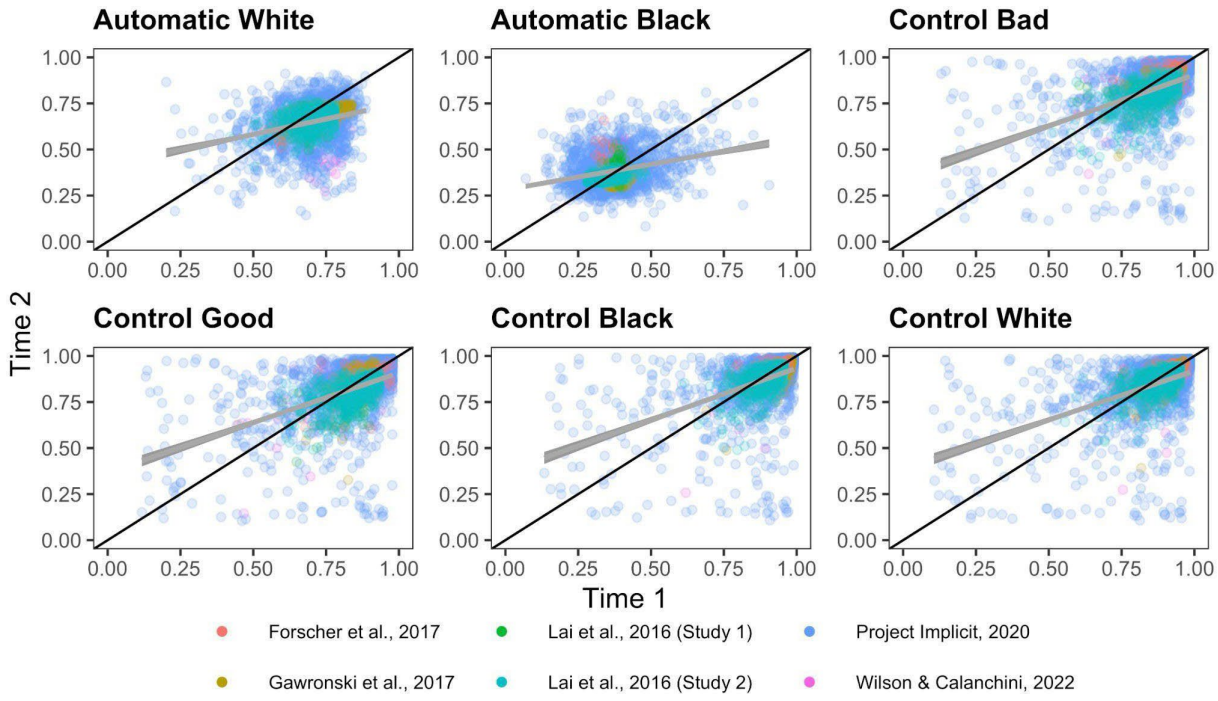

Figure 5. Scatterplots depicting the relationship between Time 1 and Time 2 PDP parameters. The black diagonal line depicts perfect consistency between Time 1 and Time 2 and the gray line depicts best-fit slopes through the observed data.

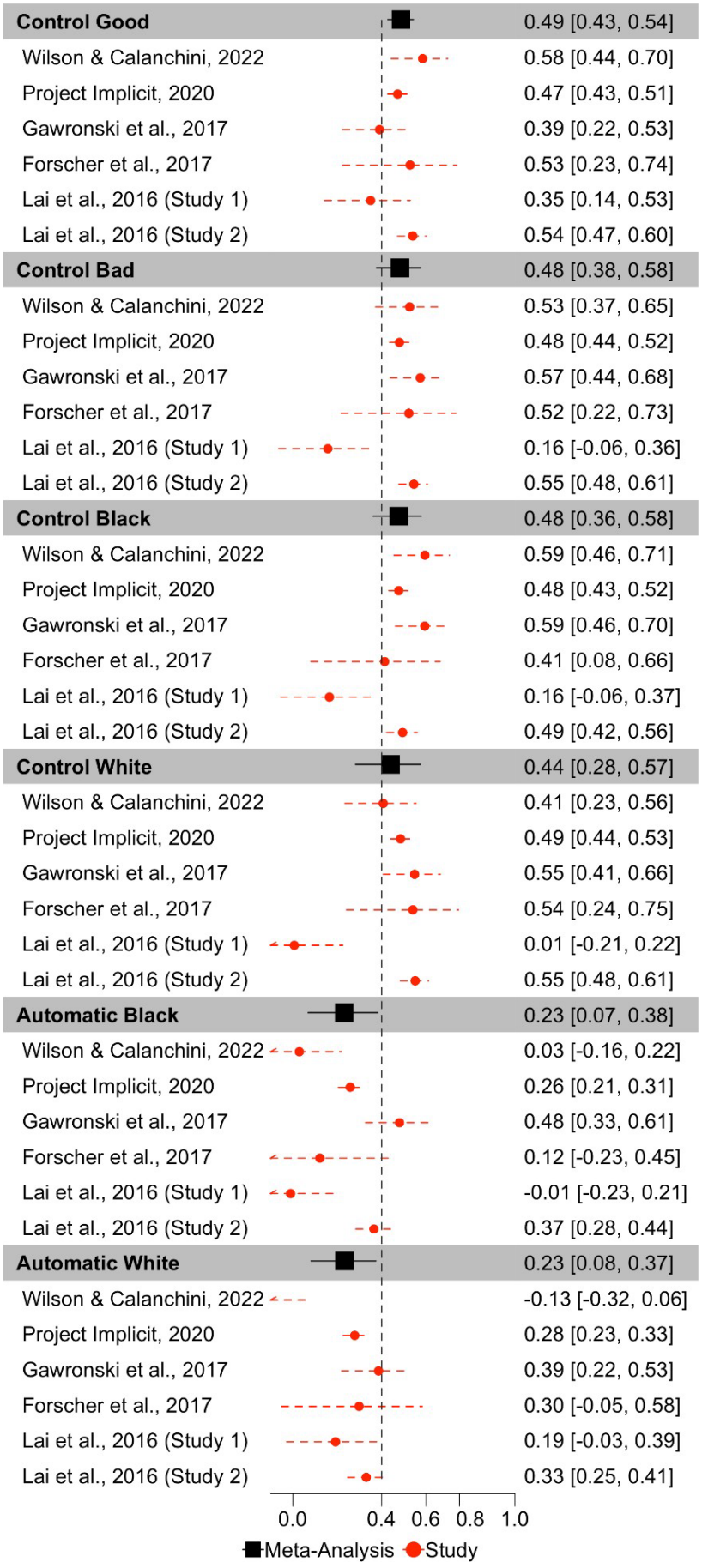
Control Good 0.49 [0.43, 0.54]
Wilson & Calanchini, 2022 0.58 [0.44, 0.70]
Project Implicit, 2020 0.47 [0.43, 0.51]
Gawronski et al., 2017 0.39 [0.22, 0.53]
Forscher et al., 2017 0.53 [0.23, 0.74]
Lai et al., 2016 (Study 1) 0.35 [0.14, 0.53]
Lai et al., 2016 (Study 2) 0.54 [0.47, 0.60]
Control Bad 0.48 [0.38, 0.58]
Wilson & Calanchini, 2022 0.53 [0.37, 0.65]
Project Implicit, 2020 0.48 [0.44, 0.52]
Gawronski et al., 2017 0.57 [0.44, 0.68]
Forscher et al., 2017 0.52 [0.22, 0.73]
Lai et al., 2016 (Study 1) 0.16 [-0.06, 0.36]
Lai et al., 2016 (Study 2) 0.55 [0.48, 0.61]
Control Black 0.48 [0.36, 0.58]
Wilson & Calanchini, 2022 0.59 [0.46, 0.71]
Project Implicit, 2020 0.48 [0.43, 0.52]
Gawronski et al., 2017 0.59 [0.46, 0.70]
Forscher et al., 2017 0.41 [0.08, 0.66]
Lai et al., 2016 (Study 1) 0.16 [-0.06, 0.37]
Lai et al., 2016 (Study 2) 0.49 [0.42, 0.56]
Control White 0.44 [0.28, 0.57]
Wilson & Calanchini, 2022 0.41 [0.23, 0.56]
Project Implicit, 2020 0.49 [0.44, 0.53]
Gawronski et al., 2017 0.55 [0.41, 0.66]
Forscher et al., 2017 0.54 [0.24, 0.75]
Lai et al., 2016 (Study 1) 0.01 [-0.21, 0.22]
Lai et al., 2016 (Study 2) 0.55 [0.48, 0.61]
Automatic Black 0.23 [0.07, 0.38]
Wilson & Calanchini, 2022 0.03 [-0.16, 0.22]
Project Implicit, 2020 0.26 [0.21, 0.31]
Gawronski et al., 2017 0.48 [0.33, 0.61]
Forscher et al., 2017 0.12 [-0.23, 0.45]
Lai et al., 2016 (Study 1) -0.01 [-0.23, 0.21]
Lai et al., 2016 (Study 2) 0.37 [0.28, 0.44]
Automatic White 0.23 [0.08, 0.37]
Wilson & Calanchini, 2022 -0.13 [-0.32, 0.06]
Project Implicit, 2020 0.28 [0.23, 0.33]
Gawronski et al., 2017 0.39 [0.22, 0.53]
Forscher et al., 2017 0.30 [-0.05, 0.58]
Lai et al., 2016 (Study 1) 0.19 [-0.03, 0.39]
Lai et al., 2016 (Study 2) 0.33 [0.25, 0.41]
0.0
0.4
0.6
0.8
1.0
Meta-Analysis
Study

Figure 6. Forest plot depicting the meta-analytic and study specific consistencies for the PDP, depicted as intraclass correlations. Point estimates and 95% confidence intervals are reported. Dashed vertical line reflects the threshold of acceptable reliability.

Parameter Recovery Within Measurement Occasions

To establish whether parameters are reliable within a measurement occasion we performed parameter recovery. These analyses provide insight into whether the data-generating process for these parameters can reliably recover the same parameters, and therefore whether the process can be reliably measured – which, in turn, illuminates the extent to which stability between measurement occasions reflects within-individual variability versus measurement error. We considered parameter recovery to be acceptable if $r > .70$ (Nunnally, 1978; Nunnally & Bernstein, 1994; Palminteri et al., 2017).

The Quad Model

Correlations between original and recovered parameters for each Quad parameter are depicted in Figure 7. Meta-analytic results (Figure 8) indicate that Detection parameters were the most recoverable of all Quad parameters, and demonstrate acceptable recovery: $r = .870$, 95% CI = [.767 - .929], $p < 001$. The other Quad parameters did not demonstrate acceptable recovery: White-good Associations $r = .590$, 95% CI = [.406 - .728], $p < .001$; Black-bad Associations $r = .588$, 95% CI = [.423 - .715], $p < .001$; Overcoming Bias $r = .259$, 95% CI = [.115 - .394], $p = .0006$; Guessing $r = .173$, 95% CI = [-.035 - .367], $p = .102$. That said, the Associations parameters both demonstrated modest recoverability that approached the threshold for acceptable, whereas Overcoming Bias and Guessing demonstrated unequivocally poor recovery that approached or included 0.

Inspection of CIs indicates that Detection exhibited significantly higher recoverability than do any of the other Quad parameters. Both Associations parameters also demonstrated

significantly higher recoverability than the Overcoming Bias and Guessing parameters. The Associations parameters did not differ from one another in terms of recoverability, nor did the Overcoming Bias and Guessing parameters differ from one another in terms of recoverability.

There was substantial heterogeneity in recoverability between-studies for Detection ($Q(11) = 600.942$, $p < .001$, $I^2 = 97.878$), Guessing ($Q(11) = 329.556$, $p < .0001$, $I^2 = 97.193$), Overcoming Bias ($Q(11) = 102.428$, $p < .0001$, $I^2 = 91.857$), White-good Associations ($Q(11) = 416.096$, $p < .0001$, $I^2 = 96.670$), and Black-bad Associations ($Q(11) = 372.995$, $p < .0001$, $I^2 = 96.603$). The amount of heterogeneity in recovery is surprising, given that recovery estimates should be relatively stable given a particular model configuration. We report exploratory analyses later in this manuscript in which we further probe this point.

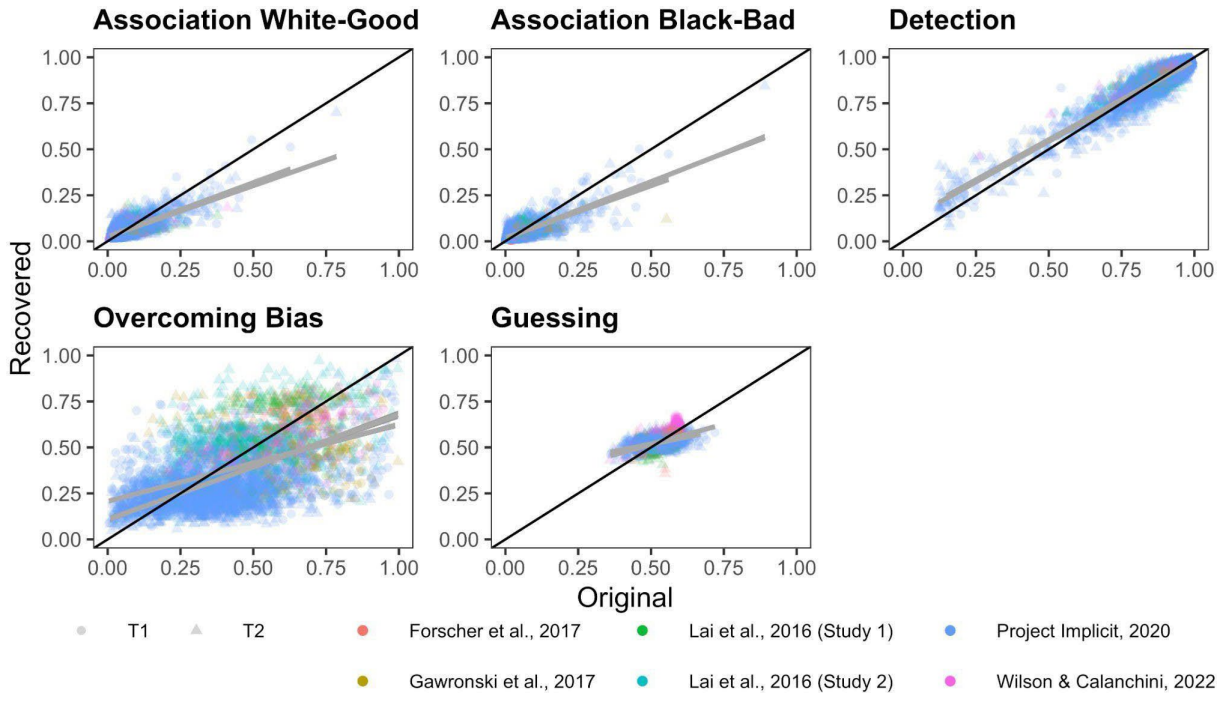

Figure 7. Scatterplots depicting the relationship between the original and simulated Quad parameters. The black diagonal line depicts perfect recovery and the gray line depicts best-fit slopes through the observed data. Each study is depicted as a different color. Time 1 and Time 2 parameter recoveries were conducted separately and are depicted as different shapes.

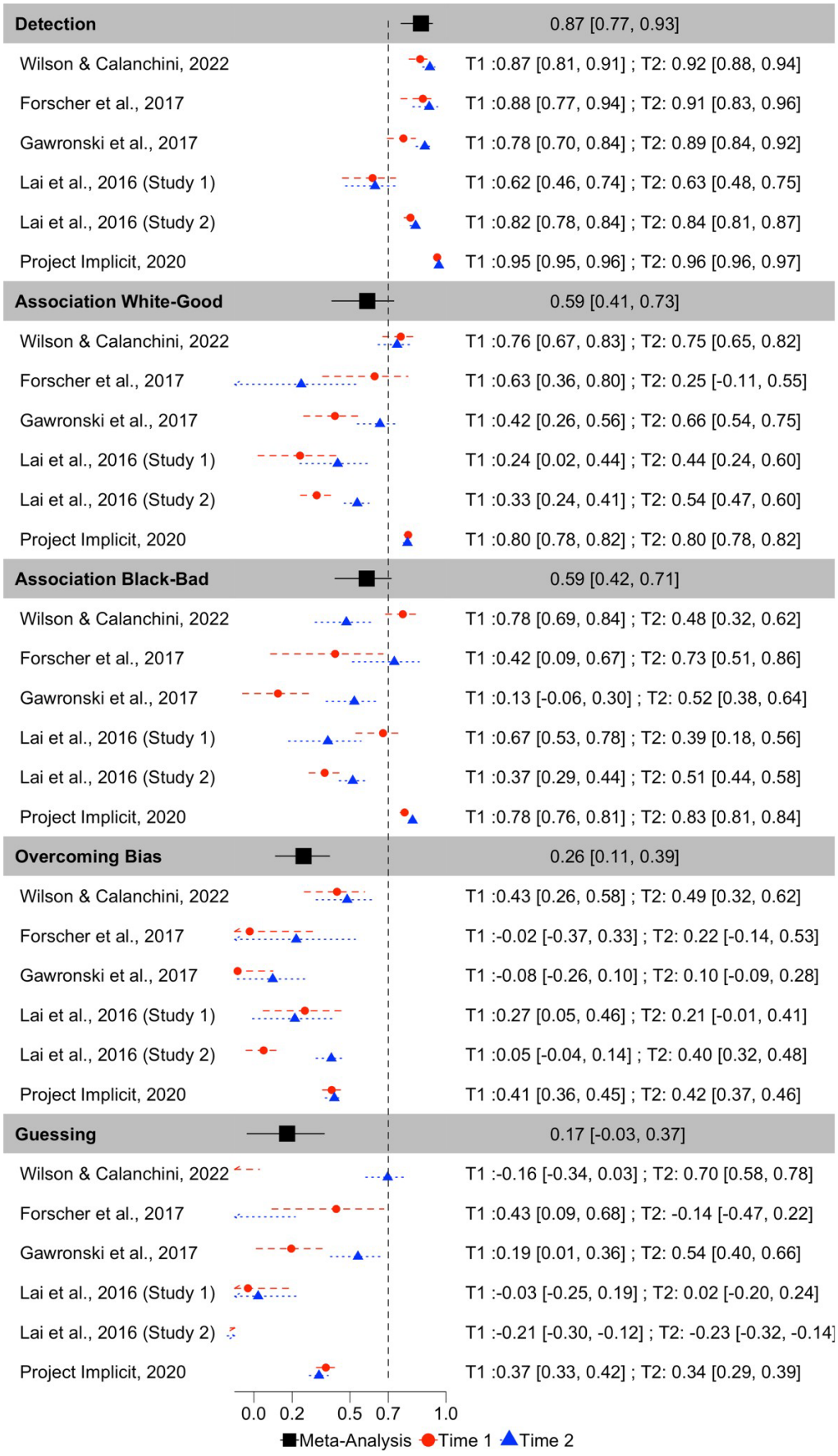
Detection 0.87 [0.77, 0.93]
Wilson & Calanchini, 2022 T1 :0.87 [0.81, 0.91] ; T2: 0.92 [0.88, 0.94]
Forscher et al., 2017 T1 :0.88 [0.77, 0.94] ; T2: 0.91 [0.83, 0.96]
Gawronski et al., 2017 T1 :0.78 [0.70, 0.84] ; T2: 0.89 [0.84, 0.92]
Lai et al., 2016 (Study 1) T1 :0.62 [0.46, 0.74] ; T2: 0.63 [0.48, 0.75]
Lai et al., 2016 (Study 2) T1 :0.82 [0.78, 0.84] ; T2: 0.84 [0.81, 0.87]
Project Implicit, 2020 T1 :0.95 [0.95, 0.96] ; T2: 0.96 [0.96, 0.97]
Association White-Good 0.59 [0.41, 0.73]
Wilson & Calanchini, 2022 T1 :0.76 [0.67, 0.83] ; T2: 0.75 [0.65, 0.82]
Forscher et al., 2017 T1 :0.63 [0.36, 0.80] ; T2: 0.25 [-0.11, 0.55]
Gawronski et al., 2017 T1 :0.42 [0.26, 0.56] ; T2: 0.66 [0.54, 0.75]
Lai et al., 2016 (Study 1) T1 :0.24 [0.02, 0.44] ; T2: 0.44 [0.24, 0.60]
Lai et al., 2016 (Study 2) T1 :0.33 [0.24, 0.41] ; T2: 0.54 [0.47, 0.60]
Project Implicit, 2020 T1 :0.80 [0.78, 0.82] ; T2: 0.80 [0.78, 0.82]
Association Black-Bad 0.59 [0.42, 0.71]
Wilson & Calanchini, 2022 T1 :0.78 [0.69, 0.84] ; T2: 0.48 [0.32, 0.62]
Forscher et al., 2017 T1 :0.42 [0.09, 0.67] ; T2: 0.73 [0.51, 0.86]
Gawronski et al., 2017 T1 :0.13 [-0.06, 0.30] ; T2: 0.52 [0.38, 0.64]
Lai et al., 2016 (Study 1) T1 :0.67 [0.53, 0.78] ; T2: 0.39 [0.18, 0.56]
Lai et al., 2016 (Study 2) T1 :0.37 [0.29, 0.44] ; T2: 0.51 [0.44, 0.58]
Project Implicit, 2020 T1 :0.78 [0.76, 0.81] ; T2: 0.83 [0.81, 0.84]
Overcoming Bias 0.26 [0.11, 0.39]
Wilson & Calanchini, 2022 T1 :0.43 [0.26, 0.58] ; T2: 0.49 [0.32, 0.62]
Forscher et al., 2017 T1 :-0.02 [-0.37, 0.33] ; T2: 0.22 [-0.14, 0.53]
Gawronski et al., 2017 T1 :-0.08 [-0.26, 0.10] ; T2: 0.10 [-0.09, 0.28]
Lai et al., 2016 (Study 1) T1 :0.27 [0.05, 0.46] ; T2: 0.21 [-0.01, 0.41]
Lai et al., 2016 (Study 2) T1 :0.05 [-0.04, 0.14] ; T2: 0.40 [0.32, 0.48]
Project Implicit, 2020 T1 :0.41 [0.36, 0.45] ; T2: 0.42 [0.37, 0.46]
Guessing 0.17 [-0.03, 0.37]
Wilson & Calanchini, 2022 T1 :-0.16 [-0.34, 0.03] ; T2: 0.70 [0.58, 0.78]
Forscher et al., 2017 T1 :0.43 [0.09, 0.68] ; T2: -0.14 [-0.47, 0.22]
Gawronski et al., 2017 T1 :0.19 [0.01, 0.36] ; T2: 0.54 [0.40, 0.66]
Lai et al., 2016 (Study 1) T1 :-0.03 [-0.25, 0.19] ; T2: 0.02 [-0.20, 0.24]
Lai et al., 2016 (Study 2) T1 :-0.21 [-0.30, -0.12] ; T2: -0.23 [-0.32, -0.14]
Project Implicit, 2020 T1 :0.37 [0.33, 0.42] ; T2: 0.34 [0.29, 0.39]
0.0 0.2 0.5 0.7 1.0
Meta-Analysis Time 1 Time 2

Figure 8. Forest plot depicting the meta-analytic and study-level (for each measurement occasion) recovery estimates for each Quad parameter, in terms of Pearson r correlations. Point estimates and 95% confidence intervals are reported. Dashed vertical line reflects the threshold of acceptable recoverability.

The Process Dissociation Procedure

Correlations between original and recovered parameters for each PDP parameter are depicted in Figure 9. Meta-analytic results (Figure 10) indicate that Control parameters were the most recoverable, with point estimates that all demonstrate acceptable recovery: Control-Black r = .816, 95% CI = [.653 - .907], p < 001; Control-bad r = .813, 95% CI = [.671 - .898], p < .001; Control-good r = .808, 95% CI = [.671 - .892], p < .001; Control-White r = .798, 95% CI = [.638 - .892], p < .001. The Automatic parameters demonstrated modest recoverability, but did not meet the a priori threshold of .70 for acceptable recoverability: Automatic-White r = .494, 95% CI = [.364 - .606], p < .0001; Automatic-Black r = .458, 95% CI = [.195 - .660], p < .0001.

Inspection of CIs indicates that all four Control parameters exhibited significantly higher recoverability than both of the Automatic parameters. The Control parameters did not differ from one another in terms of recoverability, nor did the Automatic parameters differ from one another in terms of recoverability.

There was substantial heterogeneity in recoverability between-studies for Automatic-Black (Q(11) = 236.075, p = .0001, I^2 = 97.744), Automatic-White (Q(11) = 163.855, p < .0001, I^2 = 93.730), Control-Black (Q(11) = 700.643, p < .0001, I^2 = 98.460), , Control-good (Q(11) = 630.490, p < .0001, I^2 = 97.833), Control-White (Q(11) = 598.092, p < .0001, I^2 = 98.225), and Control-bad (Q(11) = 571.550, p < .0001, I^2 = 97.893).

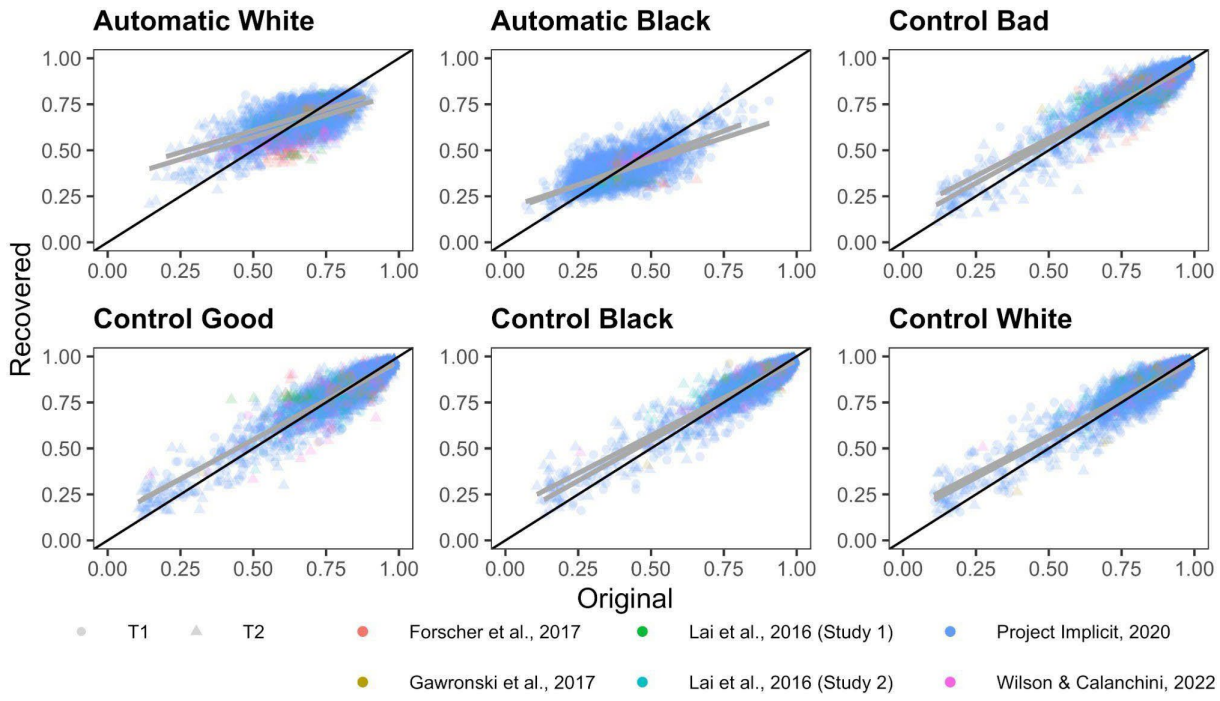

Figure 9. Scatterplots depicting the relationship between the original and simulated PDP parameters. The black diagonal line depicts perfect recovery and the gray line depicts best-fit slopes through the observed data. Each study is depicted as a different color. Time 1 and Time 2 parameter recoveries were conducted separately and are depicted as different shapes.

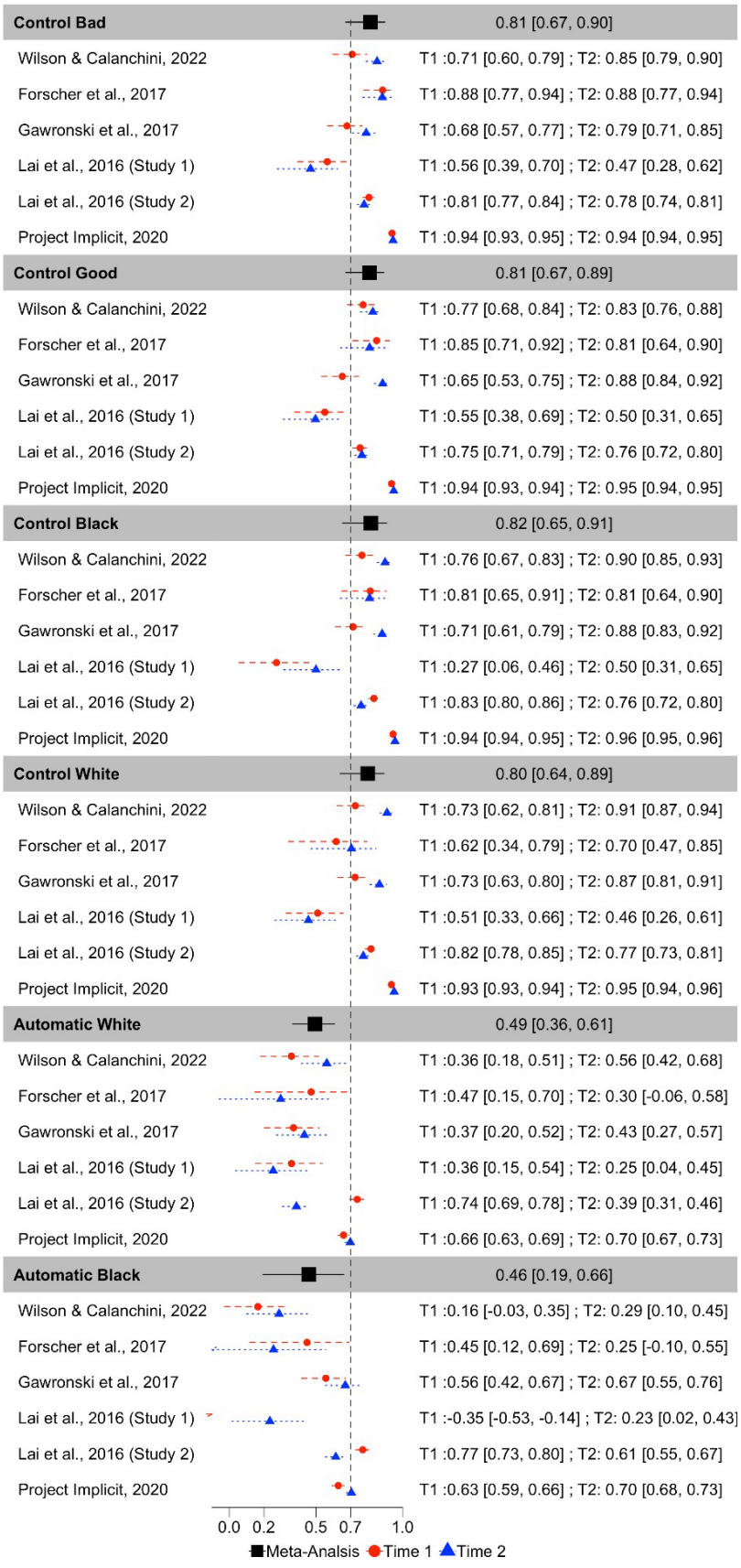
Control Bad
0.81 [0.67, 0.90]
Wilson & Calanchini, 2022
T1 :0.71 [0.60, 0.79] ; T2: 0.85 [0.79, 0.90]
Forscher et al., 2017
T1 :0.88 [0.77, 0.94] ; T2: 0.88 [0.77, 0.94]
Gawronski et al., 2017
T1 :0.68 [0.57, 0.77] ; T2: 0.79 [0.71, 0.85]
Lai et al., 2016 (Study 1)
T1 :0.56 [0.39, 0.70] ; T2: 0.47 [0.28, 0.62]
Lai et al., 2016 (Study 2)
T1 :0.81 [0.77, 0.84] ; T2: 0.78 [0.74, 0.81]
Project Implicit, 2020
T1 :0.94 [0.93, 0.95] ; T2: 0.94 [0.94, 0.95]
Control Good
0.81 [0.67, 0.89]
Wilson & Calanchini, 2022
T1 :0.77 [0.68, 0.84] ; T2: 0.83 [0.76, 0.88]
Forscher et al., 2017
T1 :0.85 [0.71, 0.92] ; T2: 0.81 [0.64, 0.90]
Gawronski et al., 2017
T1 :0.65 [0.53, 0.75] ; T2: 0.88 [0.84, 0.92]
Lai et al., 2016 (Study 1)
T1 :0.55 [0.38, 0.69] ; T2: 0.50 [0.31, 0.65]
Lai et al., 2016 (Study 2)
T1 :0.75 [0.71, 0.79] ; T2: 0.76 [0.72, 0.80]
Project Implicit, 2020
T1 :0.94 [0.93, 0.94] ; T2: 0.95 [0.94, 0.95]
Control Black
0.82 [0.65, 0.91]
Wilson & Calanchini, 2022
T1 :0.76 [0.67, 0.83] ; T2: 0.90 [0.85, 0.93]
Forscher et al., 2017
T1 :0.81 [0.65, 0.91] ; T2: 0.81 [0.64, 0.90]
Gawronski et al., 2017
T1 :0.71 [0.61, 0.79] ; T2: 0.88 [0.83, 0.92]
Lai et al., 2016 (Study 1)
T1 :0.27 [0.06, 0.46] ; T2: 0.50 [0.31, 0.65]
Lai et al., 2016 (Study 2)
T1 :0.83 [0.80, 0.86] ; T2: 0.76 [0.72, 0.80]
Project Implicit, 2020
T1 :0.94 [0.94, 0.95] ; T2: 0.96 [0.95, 0.96]
Control White
0.80 [0.64, 0.89]
Wilson & Calanchini, 2022
T1 :0.73 [0.62, 0.81] ; T2: 0.91 [0.87, 0.94]
Forscher et al., 2017
T1 :0.62 [0.34, 0.79] ; T2: 0.70 [0.47, 0.85]
Gawronski et al., 2017
T1 :0.73 [0.63, 0.80] ; T2: 0.87 [0.81, 0.91]
Lai et al., 2016 (Study 1)
T1 :0.51 [0.33, 0.66] ; T2: 0.46 [0.26, 0.61]
Lai et al., 2016 (Study 2)
T1 :0.82 [0.78, 0.85] ; T2: 0.77 [0.73, 0.81]
Project Implicit, 2020
T1 :0.93 [0.93, 0.94] ; T2: 0.95 [0.94, 0.96]
Automatic White
0.49 [0.36, 0.61]
Wilson & Calanchini, 2022
T1 :0.36 [0.18, 0.51] ; T2: 0.56 [0.42, 0.68]
Forscher et al., 2017
T1 :0.47 [0.15, 0.70] ; T2: 0.30 [-0.06, 0.58]
Gawronski et al., 2017
T1 :0.37 [0.20, 0.52] ; T2: 0.43 [0.27, 0.57]
Lai et al., 2016 (Study 1)
T1 :0.36 [0.15, 0.54] ; T2: 0.25 [0.04, 0.45]
Lai et al., 2016 (Study 2)
T1 :0.74 [0.69, 0.78] ; T2: 0.39 [0.31, 0.46]
Project Implicit, 2020
T1 :0.66 [0.63, 0.69] ; T2: 0.70 [0.67, 0.73]
Automatic Black
0.46 [0.19, 0.66]
Wilson & Calanchini, 2022
T1 :0.16 [-0.03, 0.35] ; T2: 0.29 [0.10, 0.45]
Forscher et al., 2017
T1 :0.45 [0.12, 0.69] ; T2: 0.25 [-0.10, 0.55]
Gawronski et al., 2017
T1 :0.56 [0.42, 0.67] ; T2: 0.67 [0.55, 0.76]
Lai et al., 2016 (Study 1)
T1 :-0.35 [-0.53, -0.14] ; T2: 0.23 [0.02, 0.43]
Lai et al., 2016 (Study 2)
T1 :0.77 [0.73, 0.80] ; T2: 0.61 [0.55, 0.67]
Project Implicit, 2020
T1 :0.63 [0.59, 0.66] ; T2: 0.70 [0.68, 0.73]
0.0
0.2
0.5
0.7
1.0
Meta-Analsis
Time 1
Time 2

Figure 10. Forest plot depicting the meta-analytic and study-level (for each measurement occasion) recovery estimates for each PDP parameter, in terms of Pearson r correlations. Point estimates and 95% confidence intervals are reported. Dashed vertical line reflects the threshold of acceptable recoverability.

Exploratory Analyses and Results

In addition to the pre-registered analyses reported above, we summarize below two exploratory analyses. These analyses aim to address additional questions that arose over the course of conducting the pre-registered analyses related to the repeatability of IAT responses across simulations, and to the moderating role of time between measurement occasions on parameter consistency.

Repeatability of IAT Responses Across Simulations

Given the high heterogeneity in recoverability estimates identified across samples, we investigated the repeatability with which a given set of parameters can estimate similar response frequencies. Our exploratory analyses (which are described in full in the Supplement) examined repeatability in two ways: in terms of the relationships among multiple simulations of responses, and in terms of the relationship between simulated responses and participants' original responses.

Supplemental Table 1 summarizes repeatability among simulated response frequencies for the Quad model, and Supplemental Table 2 summarizes repeatability among simulated response frequencies for the PDP. Across both models, simulated responses were weakly-to-moderately associated with other simulated responses, whereas simulated responses were more strongly associated with participants' original responses. Project Implicit is the largest sample and exhibits the strongest repeatability estimates, which potentially reflects more accurate population-level estimates that help to regularize individual-level estimates at larger sample sizes.

Time as a Moderator of Consistency Between Measurement Occasion

The amount of time elapsed between measurement occasions may reasonably be expected to moderate the extent of parameter consistency, such that shorter intervals between measurement occasions would be related to higher consistency. Based on the time intervals between measurement occasions listed in Table 1, we conducted a series of exploratory orthogonal polynomial contrasts to examine the relationship between time between measurement and consistency for each MPT parameter.

For the Quad model, there was a marginal trend of linear moderation, and a trend of quadratic moderation, on consistency by the categorical order of time interval length for Overcoming Bias, $Q_M(4) = 18.714$, $p = .001$; $\beta_{Linear} = -.239$, $SE_{Linear} = 0.126$, $p_{Linear} = 0.059$; $\beta_{Quadratic} = -.303$, $SE_{Quadratic} = .109$, $p_{Quadratic} = .006$, which suggests that Overcoming Bias may initially decrease and then increase in consistency over time. Categorical order of time length did not moderate consistency for the other Quad parameters, nor for any of the PDP parameters. Thus, there does not appear to be much evidence that time between IAT measurement occasions moderates the consistency of MPT parameters.

Discussion

The present research investigates the extent to which the processes that contribute to responses on the race IAT are stable within individuals over time and can be reliably measured. Aligning with calls for more formal modeling in psychological science (Robinaugh et al., 2021; Smaldino, 2020), we investigated these questions using the Quad model and PDP to concretely specify our theoretical assumptions, and meta-analyzed our findings across six datasets collected by multiple laboratories to increase the validity of our findings.

We found that parameters reflecting two accuracy-oriented processes (i.e., responses that correctly identify stimulus) – Detection in the Quad model and Control in the PDP – generally demonstrated fair consistency between measurement occasions and acceptable recoverability within measurement occasion. This pattern of results suggests that both parameters reflect relatively stable individual differences. In contrast, parameters reflecting associations between target groups and attributes – Associations in the Quad model and Automatic in the PDP – did not meet our a priori thresholds for fair consistency or acceptable recoverability. One interpretation of this pattern of results is that the Associations and Automatic parameters do not reflect stable individual differences. However, because these two parameters demonstrate what could be reasonably characterized as modest recoverability, with *r*s > .58 for Associations, and *r*s >.45 for Automatic, we cannot rule out the alternative possibility that the Associations and Automatic parameters reflect relatively noisily-measured individual differences. Finally, the Overcoming Bias and Guessing parameters in the Quad model both demonstrated unequivocally poor consistency and recoverability, suggesting that these parameters are unstable, likely contain a large degree of measurement error in their estimation, and should not be interpreted as individual-level estimates.

Theoretical Implications for Implicit Measures

Our findings that the accuracy-oriented processes that contribute to responses on the IAT – Detection and Control – are reliable within measurement occasions and relatively stable over time is consistent with literature suggesting that executive functions reflect stable individual differences (Beck et al., 2011; Miyake & Friedman, 2012; Paap & Sawi, 2016; Willoughby & Blair, 2011). This pattern of results also dovetails with previous research linking the Control parameter of the PDP estimated from the IAT with the executive functions of working memory

updating and task shifting (Ito et al., 2015). That said, Overcoming Bias is conceptualized as an inhibitory process, which also situates it among the constellation of executive functions – and, as such, should be expected to reflect a stable individual difference. However, the present research indicates that the Overcoming Bias parameter is neither stable nor reliable. Future research is necessary to clarify why some executive function-related parameters, such as Detection and Control, are reliable and stable, but other executive function-related parameters, such as Overcoming Bias, are unreliable and unstable.

In contrast to the pattern of results we observe for the Detection and Control parameters, our findings that Associations and Automatic parameters are relatively less stable over time would seem to pose a challenge for the perspective that responses on implicit measures reflect associations between target groups and attributes that are durably stored in memory (Greenwald et al., 1998; Petty et al., 2008). Though we cannot rule out the possibility that these parameters reflect relatively stable but noisily-measured individual differences, their low consistency and modest recoverability support a context-dependent perspective on implicit social cognition (Conrey & Smith, 2007; Payne et al., 2017; Schwarz, 2007). Context can be operationalized in a variety of ways – including physical spaces, geographical areas, social situations, internal states, or specific times – and our data can only speak to the time-dependence of model parameters. To date, much of the evidence investigating context-dependent perspectives in implicit social cognition has focused on physical space (Hannay & Payne, 2022; Ofosu et al., 2019; Vuletich & Payne, 2019). However, most of the data reflected in the present research was collected over the internet and, thus, we have little information about the physical spaces in which participants completed these IATs. Future research is needed to discern whether Associations and Automatic parameters reflect context-dependent versus stable but noisily-measured constructs (Carpenter et

al., 2022; Connor & Evers, 2020), and tightly controlled measurement settings may provide deeper insight into situational features related to the context-dependence of these constructs

We relied on two qualitatively distinct MPTs in the present research with an eye towards the validity of our findings. To the extent that we find a pattern of results across conceptually analogous parameters in each MPT, then we can have relatively high confidence that our findings do not reflect idiosyncrasies of our modeling choices. Indeed, we found a very similar pattern of results across the Detection parameter of the Quad model and the Control parameters of the PDP in terms of both consistency and recoverability. However, the Associations parameters of the Quad model were descriptively more recoverable than were the Automatic parameters of the PDP. One possible explanation for this divergence between Associations and Automatic parameters is that the Quad model includes the Guessing parameter as a catch-all, of sorts, that reflects the influence of any other processes not accounted for in the Associations, Detection, or Overcoming Bias parameters. Because the PDP does not include a Guessing parameter, Automatic parameters necessarily reflect the influence of any processes not accounted for in Control parameters. Thus, the Automatic parameters of the PDP can reasonably be conceptualized to reflect a combination of the Associations, Overcoming Bias, and Guessing parameters of the Quad model. As the present research indicates, Overcoming Bias and Guessing demonstrate unambiguously poor recoverability; consequently, their influence 'contaminates' the Automatic parameters. From this perspective, Associations parameters in the Quad model would seem to be a purer index than Automatic parameters in the PDP of the strength with which a target category is associated with an attribute. Thus, both psychometrically and theoretically, Associations parameters may be better candidates for assessing and predicting individual differences than Automatic parameters – but both exhibit only moderate reliability for

individual-level inference. Though, to be clear, we make no claims that any MPT parameters is a pure measure of any process or construct. Instead, we interpret MPT parameters to be relatively more process-pure than summary statistics (e.g., D-scores), and recognize that different parameters can vary in their process-purity.

Practical Implications for Prediction

In addition to illuminating the qualitative nature of the processes that underlie responses on implicit measures, the present research is also useful for researchers who apply formal models to their own work. Researchers often seek to correlate model parameters with theoretically-relevant individual differences measures (e.g., behavior, self-report). The reliability with which a variable can be measured imposes an upper limit on the extent to which the association between any two variables can be observed (Kanyongo et al., 2007; Loken & Gelman, 2017; Nunnally Jr., 1970; Schmidt & Hunter, 1996; Spearman, 1904). Specifically, the correlation between two measures is constrained by each measure's reliability, and thus analyses will have less statistical power if one or more variables is measured unreliably. Consequently, the Detection and Control parameters would seem to be the most promising candidates to correlate with individual difference measures because of their fair consistency between measurement occasion and acceptable recoverability within measurement occasion. Associations and Automatic parameters may be candidates to correlate with individual differences, but the extent to which their poor consistency reflects changes in "true" scores versus measurement error is unclear given their only modest parameter recoverability.

Though Associations, Automatic, Overcoming Bias, and Guessing parameters did not demonstrate acceptable recoverability in the present research, they may still be useful in some research contexts. For example, given that less reliable measures require larger samples to detect

effects, researchers who are interested in the constructs reflected in the Associations and Automatic parameters would be well-suited to rely on large datasets, such as the ones available from Project Implicit. In fact, despite low-to-moderate meta-analytic recovery estimates for these parameters, their recoverability estimates in the much larger Project Implicit datasets were generally strong: Association parameter recovery ranged .78-.83, and Automatic parameter recovery ranged .63-.70. One potential interpretation of this pattern of results is that the large samples enabled the hierarchical Bayesian estimation method to produce more accurate population estimates, which in turn produced more reliable individual-level estimates. With that said, even the more recoverable Association and Automatic parameters estimated from the Project Implicit data demonstrated poor consistency across measurement occasion (ranging .10-.14, .26-.28, respectively), which suggests that these parameters are context-dependent rather than stable but noisily measured – but future research will need to continue to investigate this point. Moreover, Overcoming Bias and Guessing recoverability estimates were poor, despite the large samples. Nevertheless, model parameters with low measurement reliability can still be robust predictors at the group-level (Hedge et al., 2018). Thus, the unequivocally poor recoverability and consistency of the Overcoming Bias parameter suggest that it is not viable for individual-level inference, but the possibility remains that it's population-level estimates may be validly examined in the context of group-level inference. Finally, Guessing demonstrated very poor psychometrics, with consistency that includes zero, so we caution against any strong interpretations of Guessing, at either the individual or group-level. Nonetheless, Guessing may still have value in model-based analyses: Guessing is configured to reflect a "catch-all", accounting for residual variance not otherwise reflected in the other model parameters, which may in turn improve the precision with which other parameters are estimated (Wilson & Collins,

2019). Indeed, the value of the Guessing parameter may be illustrated by the descriptively greater reliability of the Association parameters in the Quad model than the Activation parameters in the PDP.

Our findings would also seem to dovetail with related lines of research aimed at characterizing and predicting mental states and behavior. For example, research in functional magnetic resonance imaging (fMRI) has devoted a great deal of research to understanding test-retest reliability, with significant implications for the clinical applications of fMRI as a tool for diagnosing biomarkers of mental health risk (Bennett & Miller, 2010; Elliott et al., 2020; Herting et al., 2018). In parallel, research on economic decision-making and reinforcement learning has also interrogated the test-retest reliability of computational parameters fit to behavior (Mkrtchian et al., 2022), with potential utility for understanding mental health and psychiatric symptoms. Mirroring this work, MPT parameters can only accurately assess individual differences in the processes that contribute to responses on implicit measures if the parameters can be measured reliably. Toward that end, the present research suggests that the Detection and Control parameters are sufficiently reliable to be used to predict individual differences.

Interdisciplinary Implications for Cognitive Modeling

Given our reliance on MPT modeling, the present research is relevant to researchers across disciplines who rely on similar models rooted in the dual-process tradition of automaticity and control. Jacoby's (1991) work to disentangle the contributions of recollection and familiarity to recognition memory inspired the PDP (Payne, 2001) as we applied it in the present research. This modeling approach has also been used to investigate a wide variety of topics, including executive functioning (Ito et al., 2015), evaluative conditioning (Hütter et al., 2012), judgment and decision-making (Ferreira et al., 2006), and moral reasoning (Conway & Gawronski, 2013).

Our findings contribute to these literatures because, to our knowledge, little research has evaluated the retest reliability of MPT parameters (but see Luke & Gawronski, 2022). Similarly, our findings are relevant to researchers across disciplines who rely on response conflict-type measures like the IAT, which shares structural features with the Stroop task (Stroop, 1935), go/no-go task (Donders, 1969), and others. Models and measures like these are used across the cognitive sciences, and formal modeling offers a precisely-specified framework that can facilitate collaboration and theoretical advancement across disciplines (Calanchini et al., 2018). Consequently, the present research offers a roadmap for future investigations into the qualitative nature of a wide variety of cognitive processes.

Limitations

The present research is limited in several ways. For example, MPT modeling relies solely on response accuracy, whereas the vast majority of IAT research is based on the D-score (Greenwald et al., 2003), which relies primarily on response latency. That said, accuracy- versus latency-based operationalizations of IAT compatibility effects often reveal the same pattern of results (e.g., Meissner & Rothermund, 2013). Nevertheless, future research should explore the generalizability of our findings with modeling approaches that rely solely on response latency (Haines et al., 2020), or incorporate both response latency and accuracy (Heck & Erdfelder, 2016; Klauer et al., 2007; Klauer & Kellen, 2018).

The present research is also limited in our reliance only on the race version of the IAT. Qualitatively different cognitive processes may contribute to responses on IATs configured to assess other constructs (e.g., sexism; homophobia; stereotypes; self-concepts). That said, the Detection and Overcoming Bias parameters of the Quad model operate similarly across IATs

configured to assess different constructs (Calanchini et al., 2014), and the Control parameters of the PDP operate similarly across different implicit measures (Volpert-Esmond et al., 2020).

Relatedly, the present research is limited by its sole reliance on the IAT. The cognitive processes reflected in an implicit measure will vary depending on the structure and task demands of the measure (Payne et al., 2008). Thus, future research should investigate whether the pattern of results we report here generalize to different constructs and measures. Lastly, all of our inferences were largely based on relatively arbitrary criteria proposed by one set of researchers (Cicchetti & Sparrow, 1981), but other reasonable criteria have been proposed (Koo & Li, 2016).

Conclusion

The present research used formal modeling to investigate the reliability and stability of the processes that contribute to responses on the race IAT. Replicating across two MPTs and six independent datasets, we found that accuracy-oriented processes can be reliably measured and are somewhat stable within individuals, but other processes are less reliably measured and may vary across measurement occasions. These findings advance implicit social cognitive theory by providing insight into the temporal stability of cognitive processes that contribute to responses on implicit measures, which highlights the processes that can be expected to predict behavior and other individual differences. In turn, this work offers a model-based template for future researchers to investigate the temporal stability of cognitive processes that may be overlooked by other analytic approaches.

References

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 24–57. https://doi.org/10.1162/CPSY_a_00002

Ballard, T., Luckman, A., & Konstantinidis, E. (2020). *How meaningful are parameter estimates from models of inter-temporal choice?* https://doi.org/10.31234/osf.io/mvk67

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

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. https://doi.org/10.3758/BF03210812

Beck, D. M., Schaefer, C., Pang, K., & Carlson, S. M. (2011). Executive function in preschool children: Test–Retest reliability. *Journal of Cognition and Development*, *12*(2), 169–193. https://doi.org/10.1080/15248372.2011.563485

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*(1), 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*(3), 567–582. https://doi.org/10.1037/a0014665

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., Rivers, A. M., Klauer, K. C., & Sherman, J. W. (2018). Multinomial processing trees as theoretical bridges between cognitive and social psychology. In K. D. Federmeier (Ed.), *Psychology of Learning and Motivation* (Vol. 69, pp. 39–65). Academic Press. https://doi.org/10.1016/bs.plm.2018.09.002

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

Carpenter, T. P., Goedderz, A., & Lai, C. K. (2022). Individual differences in implicit bias can be measured reliably by administering the same Implicit Association Test multiple times. *Personality and Social Psychology Bulletin*, *0*(0). https://doi.org/10.1177/01461672221099372

Caspi, A., & Bem, D. J. (1990). Personality continuity and change across the life course. In L. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 549–575). New York, NY: Guilford Press.

Caspi, A., Bem, D. J., & Elder, G. H., Jr. (1989). Continuities and consequences of interactional styles across the life course. *Journal of Personality, 57*, 375–406.

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., Leibenluft, E., Brotman, M. A., & Cox, R. W. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping*, *39*(3), 1187–1206. https://doi.org/10.1002/hbm.23909

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater

reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345. https://doi.org/10.1177/1745691620931492

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

Conrey, F. R., & Smith, E. R. (2007). Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition*, *25*(5), 718–735. https://doi.org/10.1521/soco.2007.25.5.718

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*(2), 216–235. https://doi.org/10.1037/a0031021

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., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry*, *33*(3), 173–176. https://doi.org/10.1080/1047840X.2022.2106755

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The Mode Model as an integrative framework. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 75–109). Academic Press. https://doi.org/10.1016/S0065-2601(08)60318-4

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

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

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology*, *91*(5), 797–813. https://doi.org/10.1037/0022-3514.91.5.797

Gamer, M., Lemon, J., & Singh, I. (2010). *Irr: Various coefficients of interrater reliability and agreement*.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and

explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). *Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox*. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Hannay, J. W., & Payne, B. K. (2022). Effects of aggregation on implicit bias measurement. *Journal of Experimental Social Psychology*, *101*, 104331. https://doi.org/10.1016/j.jesp.2022.104331

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

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Herting, M. M., Gautam, P., Chen, Z., Mezher, A., & Vetter, N. C. (2018). Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies. *Developmental Cognitive Neuroscience*, *33*, 17–26. https://doi.org/10.1016/j.dcn.2017.07.001

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*(1), 116–159. https://doi.org/10.1080/10463283.2016.1212966

Hütter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating contingency awareness and conditioned attitudes: Evidence of contingency-unaware evaluative conditioning. *Journal of Experimental Psychology: General*, *141*(3), 539–557. https://doi.org/10.1037/a0026477

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

Kanyongo, G., Brook, G., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several

parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1). https://doi.org/10.22237/jmasm/1177992480

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., & 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., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, *93*(3), 353–368. https://doi.org/10.1037/0022-3514.93.3.353

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., & Wilson, M. E. (2021). Measuring implicit intergroup biases. *Social and Personality Psychology Compass*, *15*(1), e12573. https://doi.org/10.1111/spc3.12573

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Luke, D. M., & Gawronski, B. (2022). Big five personality traits and moral-dilemma Judgments: Two preregistered studies using the CNI model. *Journal of Research in Personality*, *101*,

104297. https://doi.org/10.1016/j.jrp.2022.104297

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*, *7*, e6918. https://doi.org/10.7717/peerj.6918

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

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

Melnikoff, D. E., & Bargh, J. A. (2018). The insidious number two. *Trends in Cognitive Sciences*, *22*(8), 668–669. https://doi.org/10.1016/j.tics.2018.05.005

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review*, *106*(1), 3–19.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Mkrtchian, A., Valton, V., & Roiser, J. P. (2022). Reliability of decision-making and reinforcement learning computational parameters. *BioRxiv*, 2021.06.30.450026. https://doi.org/10.1101/2021.06.30.450026

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Nunnally, J. C. (1978). An overview of psychological measurement. In B. B. Wolman (Ed.), *Clinical Diagnosis of Mental Disorders: A Handbook* (pp. 97–146). Springer US.

https://doi.org/10.1007/978-1-4684-2490-4_4

Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory*, *3*, 248–292.

Nunnally Jr., J. C. (1970). *Introduction to psychological measurement* (pp. xv, 572). McGraw-Hill.

Ofosu, E. K., Chambers, M. K., Chen, J. M., & Hehman, E. (2019). Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proceedings of the National Academy of Sciences*, *116*(18), 8846–8851. https://doi.org/10.1073/pnas.1806000116

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81–93. https://doi.org/10.1016/j.jneumeth.2016.10.002

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

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., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, *94*(1),

16–31. https://doi.org/10.1037/0022-3514.94.1.16

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/10.1080/1047840X.2017.1335568

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, *25*(5), 657–686. https://doi.org/10.1521/soco.2007.25.5.657

Petty, R. E., Fazio, R. H., & Briñol, P. (2008). *Attitudes: Insights from the new implicit measures* (pp. xix, 544). Psychology Press.

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

Roberts, B. W., Wood, D., & Caspi, A. (2008). Personality Development. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: theory and research* (3rd ed., pp. 375–398). New York, NY: Guilford Press.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Röseler, L., Wolf, D., Leder, J., & Schütz, A. (2020). *Test-Retest Reliability is not a Measure of Reliability or Stability: A Friendly Reminder*. PsyArXiv. https://doi.org/10.31234/osf.io/mt49r

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

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638–656. https://doi.org/10.1521/soco.2007.25.5.638

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smaldino, P. E. (2020). How to build a strong theoretical foundation. *Psychological Inquiry*, *31*(4), 297–301. https://doi.org/10.1080/1047840X.2020.1853463

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior.

Personality and Social Psychology Review, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(1), 1–48. https://doi.org/10.18637/jss.v036.i03

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

Vuletich, H. A., & Payne, B. K. (2019). Stability and change in implicit bias. *Psychological Science*, *30*(6), 854–862. https://doi.org/10.1177/0956797619844270

Willoughby, M., & Blair, C. (2011). Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology*, *17*(6), 564–579. https://doi.org/10.1080/09297049.2011.554390

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. https://doi.org/10.1037/0033-295X.107.1.101

Zuckerman, M. (1983). The distinction between trait and state scales is not arbitrary: Comment on Allen and Potkay's "On the arbitrary distinction between traits and states." *Journal of Personality and Social Psychology*, *44*(5), 1083–1086. https://doi.org/10.1037/0022-3514.44.5.1083

Table 1. *Spearman Rho correlations depicting the similarity of simulated IAT responses for Quad model. Each column corresponds to a different response type for simulated response frequencies. Each row corresponds to a different study and measurement occasion. The bottom row and farthest right column reflect the average of all response frequencies, and the bottom-right cell depicts the average of all correlations across studies and measurement occasions. Top table reflects the similarity among simulations. Bottom table reflects the similarity between simulations and true behavior.*

Quad Model Similarity Across Simulations

Study	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Wilson & Calanchini, 2022 T1	0.093	0.135	0.103	0.136	0.176	0.136	0.272	0.250	0.163
Wilson & Calanchini, 2022 T2	0.277	0.211	0.277	0.216	0.253	0.266	0.377	0.212	0.261
Forscher et al., 2017 T1	0.231	0.256	0.272	0.264	0.309	0.272	0.353	0.254	0.276
Forscher et al., 2017 T2	0.305	0.325	0.279	0.345	0.307	0.273	0.276	0.288	0.300
Gawronski et al., 2017 T1	0.125	0.167	0.122	0.166	0.177	0.127	0.089	0.116	0.136
Gawronski et al., 2017 T2	0.203	0.186	0.202	0.181	0.227	0.210	0.246	0.298	0.219
Lai et al., 2016 (Study 1) T1	0.046	0.072	0.048	0.077	0.113	0.102	0.159	0.187	0.100
Lai et al., 2016 (Study 1) T2	0.056	0.088	0.060	0.083	0.113	0.073	0.231	0.092	0.100
Lai et al., 2016 (Study 2) T1	0.189	0.200	0.186	0.197	0.187	0.172	0.188	0.163	0.185
Lai et al., 2016 (Study 2) T2	0.201	0.200	0.196	0.203	0.192	0.198	0.221	0.220	0.204
Project Implicit, 2020 T1	0.400	0.422	0.402	0.420	0.493	0.463	0.494	0.495	0.448
Project Implicit, 2020 T2	0.469	0.487	0.469	0.487	0.482	0.474	0.475	0.502	0.481
Average	0.216	0.229	0.218	0.231	0.252	0.231	0.282	0.256	0.239

Quad Model Similarity of Simulations to True Behavior

Study	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Wilson & Calanchini, 2022 T1	0.140	0.202	0.157	0.195	0.256	0.227	0.434	0.353	0.246
Wilson & Calanchini, 2022 T2	0.328	0.251	0.342	0.257	0.273	0.334	0.440	0.292	0.315
Forscher et al., 2017 T1	0.298	0.273	0.314	0.327	0.447	0.262	0.486	0.340	0.343
Forscher et al., 2017 T2	0.372	0.378	0.318	0.489	0.363	0.259	0.463	0.401	0.380
Gawronski et al., 2017 T1	0.221	0.145	0.128	0.287	0.276	0.191	0.221	0.237	0.213
Gawronski et al., 2017 T2	0.229	0.232	0.309	0.236	0.350	0.258	0.368	0.477	0.307
Lai et al., 2016 (Study 1) T1	0.113	0.094	0.079	0.162	0.186	0.165	0.288	0.329	0.177
Lai et al., 2016 (Study 1) T2	0.114	0.158	0.132	0.192	0.167	0.149	0.414	0.226	0.194
Lai et al., 2016 (Study 2) T1	0.235	0.225	0.249	0.241	0.258	0.220	0.305	0.244	0.247
Lai et al., 2016 (Study 2) T2	0.225	0.231	0.252	0.268	0.261	0.264	0.366	0.342	0.276
Project Implicit, 2020 T1	0.450	0.459	0.439	0.491	0.544	0.506	0.552	0.551	0.499
Project Implicit, 2020 T2	0.515	0.506	0.520	0.541	0.527	0.485	0.517	0.540	0.519
Average	0.270	0.263	0.270	0.307	0.326	0.277	0.404	0.361	0.310

Table 2. *Spearman Rho correlations depicting the similarity of simulated IAT responses for PDP. Each column corresponds to a different response type for simulated response frequencies. Each row corresponds to a different study and measurement occasion. The bottom row and farthest right column reflect the average of all response frequencies, and the bottom-right cell depicts the average of all correlations across studies and measurement occasions. Top table reflects the similarity among simulations. Bottom table reflects the similarity between simulations and true behavior.*

PD Model Similarity Across Simulations

Study	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Wilson & Calanchini, 2022 T1	0.093	0.135	0.103	0.136	0.176	0.136	0.272	0.250	0.163
Wilson & Calanchini, 2022 T2	0.277	0.211	0.277	0.216	0.253	0.266	0.377	0.212	0.261
Forscher et al., 2017 T1	0.231	0.256	0.272	0.264	0.309	0.272	0.353	0.254	0.276
Forscher et al., 2017 T2	0.305	0.325	0.279	0.345	0.307	0.273	0.276	0.288	0.300
Gawronski et al., 2017 T1	0.125	0.167	0.122	0.166	0.177	0.127	0.089	0.116	0.136
Gawronski et al., 2017 T2	0.203	0.186	0.202	0.181	0.227	0.210	0.246	0.298	0.219
Lai et al., 2016 (Study 1) T1	0.046	0.072	0.048	0.077	0.113	0.102	0.159	0.187	0.100
Lai et al., 2016 (Study 1) T2	0.056	0.088	0.060	0.083	0.113	0.073	0.231	0.092	0.100
Lai et al., 2016 (Study 2) T1	0.189	0.200	0.186	0.197	0.187	0.172	0.188	0.163	0.185
Lai et al., 2016 (Study 2) T2	0.201	0.200	0.196	0.203	0.192	0.198	0.221	0.220	0.204
Project Implicit, 2020 T1	0.400	0.422	0.402	0.420	0.493	0.463	0.494	0.495	0.448
Project Implicit, 2020 T2	0.469	0.487	0.469	0.487	0.482	0.474	0.475	0.502	0.481
Average	0.216	0.229	0.218	0.231	0.252	0.231	0.282	0.256	0.239

PD Model Similarity of Simulations to True Behavior

Study	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Wilson & Calanchini, 2022 T1	0.154	0.254	0.258	0.258	0.279	0.320	0.506	0.384	0.302
Wilson & Calanchini, 2022 T2	0.392	0.313	0.447	0.368	0.437	0.490	0.466	0.368	0.410
Forscher et al., 2017 T1	0.306	0.330	0.437	0.333	0.466	0.429	0.593	0.457	0.419
Forscher et al., 2017 T2	0.260	0.390	0.400	0.557	0.247	0.281	0.532	0.497	0.395
Gawronski et al., 2017 T1	0.306	0.248	0.163	0.280	0.284	0.308	0.251	0.312	0.269
Gawronski et al., 2017 T2	0.262	0.232	0.320	0.212	0.395	0.272	0.421	0.374	0.311
Lai et al., 2016 (Study 1) T1	0.162	0.042	0.120	0.224	0.287	0.089	0.306	0.429	0.208
Lai et al., 2016 (Study 1) T2	0.226	0.161	0.243	0.231	0.234	0.190	0.407	0.254	0.243
Lai et al., 2016 (Study 2) T1	0.322	0.224	0.291	0.270	0.280	0.248	0.352	0.293	0.285
Lai et al., 2016 (Study 2) T2	0.311	0.263	0.255	0.281	0.287	0.381	0.361	0.356	0.312
Project Implicit, 2020 T1	0.512	0.488	0.490	0.530	0.638	0.584	0.584	0.569	0.549
Project Implicit, 2020 T2	0.544	0.547	0.574	0.589	0.617	0.566	0.571	0.566	0.572
Average	0.313	0.291	0.333	0.344	0.371	0.347	0.446	0.405	0.356

Reliability of IAT Responses Across Simulations

Analysis Plan

We performed 200 simulations of participant responses given their set of Quad model and PDP parameters. For both models, IAT responses are represented as a vector of 8 different response frequencies for each participant[1]. We performed two separate tests for each model. First, we examined how similar the simulated response frequencies were with one another across all simulations. For each study, we extracted the *n*th response type (from 1 to 8) for all 200 simulations and, because these responses consist of non-parametric count data, we correlated them with each other using Spearman Rho. The resulting correlation matrix reflects the correlations for the *n*th response type across 200 simulations for study *i*. We estimated the mean of the lower triangle of the correlation matrix. We iterated this process across each response type and each study, generating a table of "similarity" estimates between simulated response frequencies of each response type and within each study. Finally, we estimated the mean of each response type's similarity estimates within each study, as well as the mean of all studies' similarities within each response type, in order to provide insight into how consistent response frequencies are simulated given each participant's original MPT parameters.

As a second, complementary test, we examined how similar the simulated response frequencies were with participants' true response frequencies. This test operated similarly to the first test, except that we iterated through studies and response types, then estimated the Spearman Rho correlation of each simulation's response type *n* with the original study's response type *n*. We calculated these correlations for all simulations within a study and response type, then

[1] The IAT consists of 16 response categories: correct and incorrect responses to Black, White, good, and bad stimuli in compatible and incompatible blocks. However, incorrect responses are the complement to correct responses and, thus, they are redundant to one another. Therefore, we only modeled correct responses in this analysis.

averaged the correlations together for study *i* and count type *n*, iteratively across all studies and response types. This process produced a table of "similarity" estimates, which provides insight into how similar the response frequencies simulated from participants' parameters were to participants' original response frequencies.

Time as a Moderator of Retest Reliability

Analysis Plan

Approximate time intervals between measurement occasions are listed in Table 1, but we do not have precise information about measurement intervals for all datasets. Specifically, in the context of the Project Implicit dataset (which is the largest dataset by an order of magnitude), we know that participants completed both IATs within the same browser session, but have no information about how much time elapsed between measurements. Consequently, we assumed that browser sessions are relatively short on average, and ordinally ranked the datasets as follows for exploratory analysis:

1. Project Implicit, 2020
2. Wilson & Calanchini, 2022
3. Lai et al., 2016 (Study 1)
3. Lai et al., 2016 (Study 2)
4. Gawronski et al., 2017
5. Forscher et al., 2017

In this ranking, 1 reflects the shortest interval between measurement occasions (i.e., one browser session), and 5 reflects the longest interval between measurement occasions (i.e., 2 years). We treated measurement interval as an ordered categorical factor and modeled the interval moderator as an orthogonal polynomial contrast.

Comparing Test-Rest Reliability and Parameter Recovery

We investigated whether parameters differ in their within-measurement recoverability and between-measurement reliability.

Analysis Plan

We examined the extent to which recovery rates and retest reliability differed within each parameter by inspecting overlapping confidence intervals between test-retest ICCs and recovery correlations. These analyses provide exploratory insight into the extent to which a parameter's recoverability aligns with its stability across time.

Results

The Detection and Black-bad Associations parameters of the Quad model are significantly more recoverable within measurement occasions than they are reliable across measurement occasions. However, the White-good Associations, Overcoming Bias, and Guessing parameters do not differ in their recoverability versus retest reliability.

All four Control parameters of the PDP, along with the Automatic-Black parameter, are significantly more recoverable within measurement occasions than they are reliable across measurement occasions. However, the Automatic-White parameter is more reliable across measurement occasions than it is recoverable within measurement occasions.