A Method for Estimating the Time of Initiating Correct Categorization in Mouse-Tracking

David S. March[1] and Lowell Gaertner[2]

[1]Department of Psychology, Florida State University

[2]Department of Psychology, University of Tennessee

Corresponding Author: David S. March; march@psy.fsu.edu; Florida State University, 1107 W. Call St., Tallahassee, FL, 32304

Author Note: We thank Jonathan Freeman and Eric Hehman for graciously sharing data from Freeman et al. (2010) and Hehman et al. (2014).

To appear as:

March, D. S., & Gaertner, L. (2021). A method for estimating the time of initiating correct categorization in mouse-tracking. *Behavior Research Methods*.

Abstract

Mouse-tracking facilitates exploration of the mental processes underlying decision making. As the cognitive system works to settle on a decision, response competition manifests in the motor movements of the hand, bringing the mouse relatively closer to one alternative versus the other. Many metrics provide insight into decision-making processes by indexing the shape or complexity of the mouse trajectory. Lacking, however, is a metric that estimates the point in time when a participant begins to correctly categorize a stimulus. We rectify this absence by introducing a metric we refer to as *time of initiating correct categorization* (TICC), which is the point in time when people began moving relatively closer to the selected target relative to the distractor. We briefly review existing approaches to measuring time in mouse-tracking before describing the TICC and demonstrating its utility in three data sets.

Mouse-tracking is a process-tracing method that indexes cognitive processes online and in real-time. The temporal precision of mouse-tracking facilitates exploration of mental processes underlying decision making (i.e., categorization; Freeman & Ambady, 2010; Schoemann et al., 2019). Imagine a computer displaying a stimulus (e.g., an image of an emotionally-neutral man) at the bottom-center of the screen and two response labels at the screen's top left and right corners – one label, the target, describes the stimulus (e.g., Calm) and the other label, the distractor, does not (e.g., Dangerous). Upon stimulus presentation, the cognitive system begins to gradually accrue evidence in favor of one category versus another (i.e., response competition; Freeman et al., 2011). Mouse-tracking records the x-, y-coordinates of the mouse cursor in time as the participant moves the mouse from the bottom-center of the screen to the target. As the cognitive system works to settle on a decision, response competition manifests in the motor movements of the hand, bringing the mouse relatively closer to one alternative versus the other (Freeman, 2018; Spivey et al., 2005). For example, effects of stimulus typicality on judgement yield mouse-movement toward "Fish" before arriving at "Mammal" when categorizing a whale (Dale et al., 2007), and toward "Male" before arriving at "Female" when categorizing a masculine woman (Freeman et al., 2008).

As a process-tracing method, mouse-tracking (and other-process tracing methods, e.g., eye-tracking, electroencephalography) allows the observation of cognitive dynamics in action and yields temporal information about the development of a response. In mouse-tracking studies, the process being traced is often stimulus categorization (i.e., decision-making, judgment), and information about when categorization begins is highly relevant (Oppenheimer & Kelso, 2015). Many existing mouse-tracking metrics provide insight into decision-making processes by indexing the shape or complexity of the mouse trajectory, or how long it took to complete a trial

(Freeman & Ambady, 2010; Hehman et al., 2015). Lacking, however, is a metric that estimates the point in time when a participant *begins* to correctly categorize a stimulus in regard to the target label. That is, the time when the cognitive system begins to settle on the correct decision to a degree sufficient to result in motor movement relatively closer to the target. We rectify this absence by introducing a metric we refer to as the *time of initiating correct categorization* (TICC).[1] We briefly review existing approaches to measuring time in mouse-tracking before describing a method for estimating TICC and demonstrating its utility in three data sets.

Measuring Time in Mouse-Tracking

Depending on hardware, software, and user settings, the x-,y- position of the mouse is recorded at a given resolution in time. For example, with a mouse that samples at 100 Hz and MouseTracker software (Freeman et al., 2010), the x-,y- position is recorded every 10 ms. To manage what might be hundreds of recordings in a single trial, downsampling of raw time data can average multiple coordinates across a specified time span (e.g., downsampling in 20 ms bins would average x-,y- coordinates in adjacent 10 ms samples). Time can be modeled in raw units or normed to adjust for time differences across trials. For example, a trial lasting 1000 ms will have 100 recorded coordinates (50 if downsampled to 20 ms), whereas a trial lasting 1500 ms will have 150 recorded coordinates (75 if downsampled to 20 ms). Norming divides all trials into an equivalent number of time-steps (e.g., 101 steps; Spivey et al., 2005) and retracts or extends the measured coordinates across the steps. For example, norming a 1000 ms trial and 1500 ms trial would distort timing such that an early movement in a 1000 ms trial could be extended into a later time-step than an equally early movement in a 1500 ms trial. Because of such distortion, norming can mask real effects and create phantom effects (Gallivan & Chapman, 2014) and we

[1] In this document, measurements of time are stimulus-locked (i.e., in regard to stimulus onset).

recommend the use of raw-time when time is of interest.

There are two basic time metrics, but neither effectively assesses the time correct categorization begins, and there is an existing approach that attempts to capture the latter. Response time (RT) assesses the total duration of a trial, typically calculated as the difference in time from trial start (i.e., stimulus onset) to trial conclusion. It could be argued that RT will be shorter when correct categorization begins earlier. However, that is not always the case due to changes in acceleration and velocity across the trial (Hehman et al., 2015). In any event, RT does not directly assess the time correct categorization initiated, only when the final click occurred.

Time of maximum deviation (MD-time) is the point in time when the mouse most deviates from a direct trajectory to the target. Figure 1a displays a hypothetical trial in which the participant clicked a “Start” button to begin the trial with the display of a stimulus (i.e., an emotionally neutral face) and smoothly moved the mouse to categorize the face as “Calm” (target) or “Dangerous” (distractor). Estimation begins by computing the direct straight-line trajectory from start to target (black line in Figure 1) as a reference against the actual trajectory (blue line in Figure 1). The MD (dotted line in Figure 1) is the largest perpendicular deviation of the actual trajectory from the direct trajectory and MD-time is the time that deviation occurred. It might be tempting to interpret MD-time as when correct categorization initiated. Indeed, the actual trajectory in Figure 1a turned toward the target near the MD. However, consider the actual trajectory in Figure 1b, which is a trial from March et al. (2021). The MD (dotted line) occurred later than the next largest deviation (dashed line) and substantially later than when the actual trajectory turned toward the target. MD-time, like RT, does not directly assess the time correct categorization initiated.

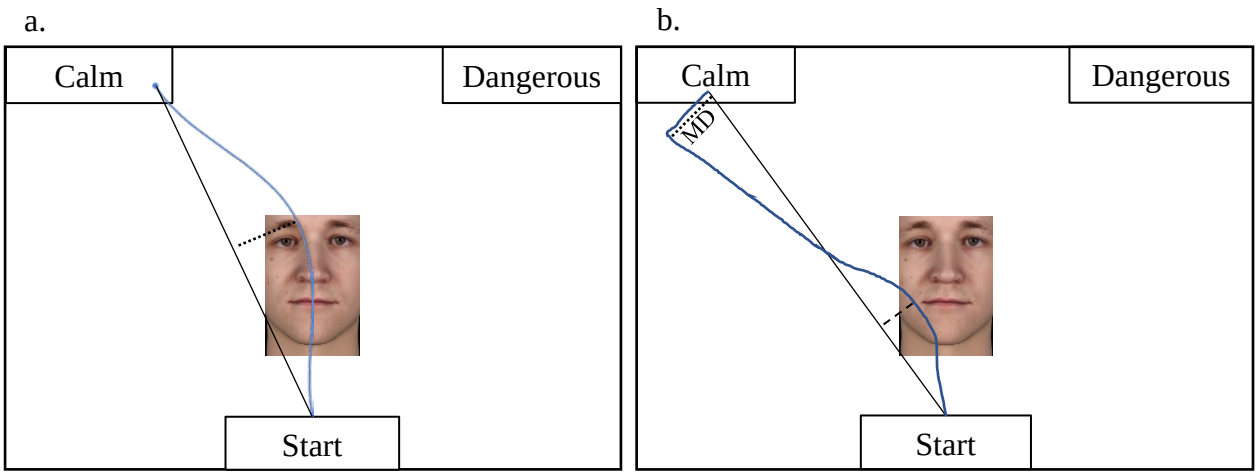

Figure 1. Actual trajectory (blue line), direct trajectory (black line), maximum deviation (dotted line), and next largest deviation (dashed line).

Spivey et al. (2005) developed an approach to assess when in time participants begin correctly categorizing a stimulus. They visually presented target and distractor images that were phonologically similar (e.g., candy and candle) or dissimilar (e.g., candy and jacket) and aurally presented via a sound file the stimulus (e.g., candy). Participants moved the mouse from a start button to an image to categorize the stimulus. To test whether correct categorization began earlier in the dissimilar than similar condition, Spivey et al. created an average trajectory for each condition time-normed into 101 time-steps and calculated at each time-step the Euclidean proximity (i.e., proportional distance in horizontal and vertical space) to the target and to the distractor. To determine when the trajectory began moving increasingly closer to the target than distractor, they performed at every time-step a dependent-samples t-test contrasting Euclidian proximity to the target vs. distractor. Proximities began to differ significantly at the 43rd time-step in the dissimilar condition and at the 76th time-step in the similar condition suggesting that phonological similarity delayed the onset of correct categorization. Subsequent studies have used this approach with variation in computing Euclidean proximity (Freeman et al., 2010), Euclidean distance (Farmer et al., 2007), horizontal distance (Dale et al., 2007; Hehman et al., 2014), or

regression (beta) weights of the relationship between an x- location or x- y-angle at each time point (Gallivan & Chapman, 2014; Sullivan et al. 2015; Scherbaum & Dschemuchadse, 2020).

Regardless of the distance or relationship computation, this approach has a limitation that incurs two problems. Time (actual or normed) is not directly estimated. Instead, time is inferred from the patterning of multiple null-hypothesis tests across time-points. In Spivey et al.'s dissimilar condition, the first 42 tests were non-significant implying that at each of those 42 time-steps the trajectory was no closer to the target than distractor (i.e., correct categorization had not initiated). One problem of inferring time from null-hypothesis tests is that the time inference is influenced, in part, by sample size. A study with a larger sample would have greater statistical power to detect less difference in distance to target vs. distractor and, therefore infer an earlier initiation time than would the same study with a smaller sample. A second problem is that in the absence of multiple trials utilizing the same stimulus, time cannot be inferred at the person-level. This hinders exploration of individual differences in time as a cause, consequence, or covariate of other phenomena linked to the person (for elaboration on these problems see the Supplemental Information). Our method for estimating TICC builds on the logic of Spivey et al.'s approach. But it directly estimates time and does not incur the latter problems.

A Method for Estimating the Time of Initiating Correct Categorization

Estimating TICC begins by averaging all trajectories for a given stimulus for a given participant – it can be done on the trial level, but an average trajectory is certainly more stable. We recommend using raw-time trajectories to maintain the fidelity of x-, y- coordinates in time (but normed-time trajectories could be used). For a given trajectory (average or single trial), the Euclidean distance (or proximity) to the target and to the distractor is calculated at each time

point.[2] Next, a difference is calculated at each time point by subtracting the Euclidean distance to the target (taken here as the last x-, y- coordinate) from the Euclidean distance to the distractor (these differences are what Spivey et al. assess with t-tests). Plotting the difference over time reveals a sigmoid curve, as is displayed in Figure 2. The flat part of the curve early in time is vertical movement bringing the mouse equally close to the target and distractor. The exponential slope is movement relatively closer to the target and further from the distractor, which flattens later in time as the mouse reaches the target.

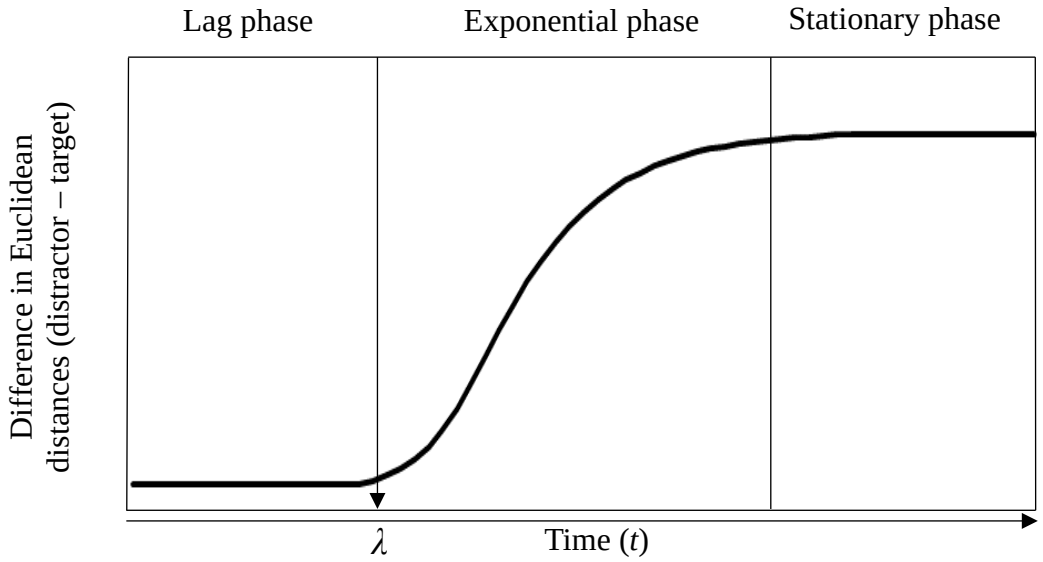

Figure 2. Difference of Euclidean distances (distractor – target) over time.

A sigmoid over time occurs for many phenomena, such as bacterial growth. The phases

[2]Euclidean distance (or proximity) is superior to only vertical or horizontal distance. Vertical movement brings the trajectory closer to the target and distractor and horizontal movement lower on the screen is further from the target (and distractor) than is the same horizontal movement higher on the screen. Euclidean distance simultaneously accounts for vertical and horizontal distance:

$$d_i = \sqrt{(x_i - x_f)^2 + (y_i - y_f)^2}$$

In the formula, d_i is distance at a given time, x_i and y_i is the horizontal and vertical location at a given time, and x_f and y_f is the horizontal and vertical location at the final time (i.e., location when the participant clicked the target – or, for distance from distractor, the corresponding location in the distractor).

on top of Figure 2 are what bacteriologists refer to as the lag phase in which growth is dormant, the exponential phase in which growth multiplies, and the stationary phase in which growth has maximized. Lambda (λ) on the time-axis is, for bacteriologists, the time when bacteria transition from dormancy to exponential growth. For the plotted difference in Euclidean distances, λ is the time when the mouse trajectory begins moving increasingly closer to the target than distractor; that is, λ is the TICC.

A number of nonlinear models estimate λ and the other the parameters of the sigmoid. Two of those models, Gompertz (Borglin et al., 2012) and Baranyi (Baty & Delignette-Muller, 2004), are highly reliable and accurate (Baty & Delignette-Muller, 2004):

$$y_t = y_{min} + y_{max} e\left\{-e\left[\frac{\mu_m e}{y_{max}}(\lambda - t) + 1\right]\right\} \text{ Gompertz}$$

$$y_t = y_{max} + \ln\left(\frac{-1 + e^{\mu_m \lambda} + e^{\mu_m t}}{(-1 + e^{\mu_m t}) + e^{(\mu_m \lambda + y_{max} - y_{min})}}\right) \text{ Baranyi}$$

In terms of mouse trajectory, y_t is the difference in Euclidean-distances at a given time (t), y_{min} is the lower asymptote of the difference, y_{max} is the upper asymptote of the difference, μ_m is the maximum growth rate, e is a mathematical constant ≈ 2.718 (i.e., Euler's number), and λ is our parameter of interest – that is, TICC, time when the trajectory begins moving increasingly closer to the target than distractor.

Software capable of nonlinear regression can estimate the parameters of the Gompertz and Baranyi models from each participant's Euclidean-distance difference at each time point. We use Proc NLIN of SAS, which requests starting values to facilitate the iterative estimation of model parameters. Starting values can be roughly guessed by eye-balling a plot of the average curve over time (i.e., averaging across all curves). Proc NLIN allows boundaries for parameter estimates. For example, λ (i.e., TICC) can be restricted to be no lower than 0 (i.e., categorization

cannot begin earlier than the trial) and no greater than the maximum trial duration, and y_{max} can be restricted within the lowest and highest possible difference in Euclidean distances (e.g., 0 and 2 if using MouseTracker; Freeman & Ambady, 2010). Although boundaries are not necessary, they yield, in our experience (detailed in the next section), higher rates of model convergence (exceeding 97%). Maximizing convergence is desirable because nonconvergence yields a missing TICC estimate. Gompertz and Baranyi models provide exceptional fit to the data, again in our experience, with average pseudo-R^2 exceeding .94 (we discuss how well TICC estimation works for different trajectory shapes in the "Applicability to Differently Shaped Trajectories" section). With no reason to prefer the TICC from the Gompertz versus Baranyi model, we average them. Hence, this procedure provides a person (or trial) specific estimate of the time correct categorization initiated for a given stimulus. That estimate can be used as a predictor or an outcome. We subsequently demonstrate the utility of TICC in three studies.

Applying TICC to Three Mouse-Tracking Studies

We demonstrate TICC from one of our studies (March et al., 2021, Study 3), and two previously published studies (Freeman et al., 2010 Study 1; Hehman, et al., 2014, Study 1). We present our study because we developed TICC for that study to test an a priori hypothesis. We present the other studies to demonstrate what could have been done had TICC been available.

TICC in March et al. (2021, Study 3)

Participants completed six mouse-tracking tasks in a study testing whether White Americans uniquely associate Black Americans as a survival threat. To demonstrate TICC, we describe one of the tasks (N =117 of 118, one person exceeded the 2 s limit on every trial and lacked recorded data). Participants clicked a start button to display a centrally-located face that was Asian, Black, or White and angry or neutral in expression, and categorized the face by

moving the mouse to click “Dangerous” or “Calm.” Dangerous was the target for angry faces and the distractor for neutral faces. Calm was the target for neutral faces and the distractor for angry faces. There were 60 such trials with 10 trials for each Race x Expression pair. Upon clicking the Start button, participants had 300 ms to initiate movement otherwise a message encouraging faster initial movement would display after the trial ended. Remaining trials ended after a decision was made via a mouse click or 2000 ms elapsed (see Figure3). To facilitate moving while deciding, rather than moving after deciding, participants completed 10 practice trials categorizing pictures of fruits and vegetables as Fruit or Vegetable. SAS code, data, and a tutorial of the subsequent steps can be found at https://osf.io/67xzv/.

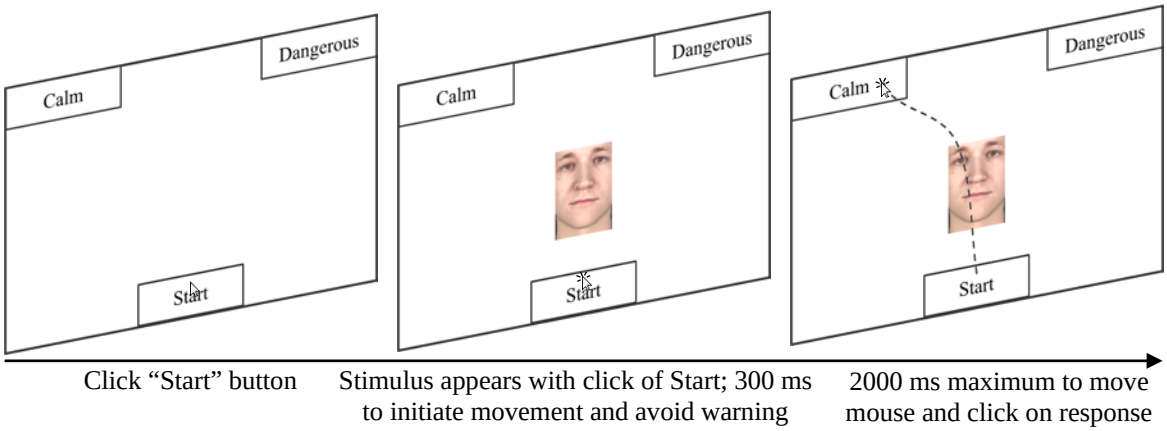

Figure 3. Trial time course.

Across the 7,020 trials, participants exceeded the 2000 ms limit from stimulus onset on 44 trials, responded incorrectly on 75 trials (clicked the distractor), initiated movement after 300 ms on 470 trials, or finished the trial under 600 ms on 204 trials, yielding 6,227 useable trials. (Conclusions based on inferential tests and direction of effects are unchanged when excluded trials are retained). March et al. (a) created an average mouse-trajectory for each participant for each Race x Expression pair (702 average trajectories), (b) calculated at each (raw) time point the difference in Euclidean distance to target and distractor for each trajectory, (c) estimated the

latter sigmoid with the Gompertz and Baranyi models in SAS Proc NLIN, which converged on 701 of 702 trajectories and evidenced exceptional fit with average pseudo-R^2 of .9499 and .9522, respectively, and (d) averaged the estimates from each model yielding six TICCs per participant (i.e., one TICC for each Race x Expression pair). Figure 4 displays for each race by expression pair the average mouse trajectory (top panel), difference in Euclidian distance over time (middle panel), and estimated difference in Euclidian distance over time from the averaged Gompertz and Baranyi estimates with area of focus on TICC (grey area and insert in bottom panel).

A repeated measures ANOVA on TICC scores was consistent with the hypothesis that White Americans associate Black (more than White or Asian) Americans as a survival threat. In particular, there was a Race x Expression interaction, $F(2, 114) = 24.76$, $p < .0001$, such that participants began categorizing angry faces as dangerous *earlier* in time when the faces were Black (M = 496 ms, SD = 75) than White (M = 527 ms, SD = 75), $F(1, 115) = 19.21$, $p = .0001$, $d = 0.41$, or Asian (M = 529 ms, SD = 115), $F(1, 115) = 8.37$, $p = .0046$, $d = 0.27$, and the latter two did not differ, $F(1, 115) = 0.02$, $p = .9018$, $d = 0.01$, and they began categorizing neutral faces as calm *later* in time when the faces were Black (M = 566 ms, SD = 101) than White (M = 517 ms, SD = 80), $F(1, 115) = 23.47$, $p = .0001$, $d = 0.45$, or Asian (M = 524 ms, SD = 111), $F(1, 115) = 16.38$, $p = .0001$, $d = 0.38$, and the latter two did not differ, $F(1, 115) = 0.31$, $p = .5771$, $d = 0.05$. By estimating time with TICC, March et al. were able to efficiently test their hypothesis regarding the role of threat in anti-Black prejudice.

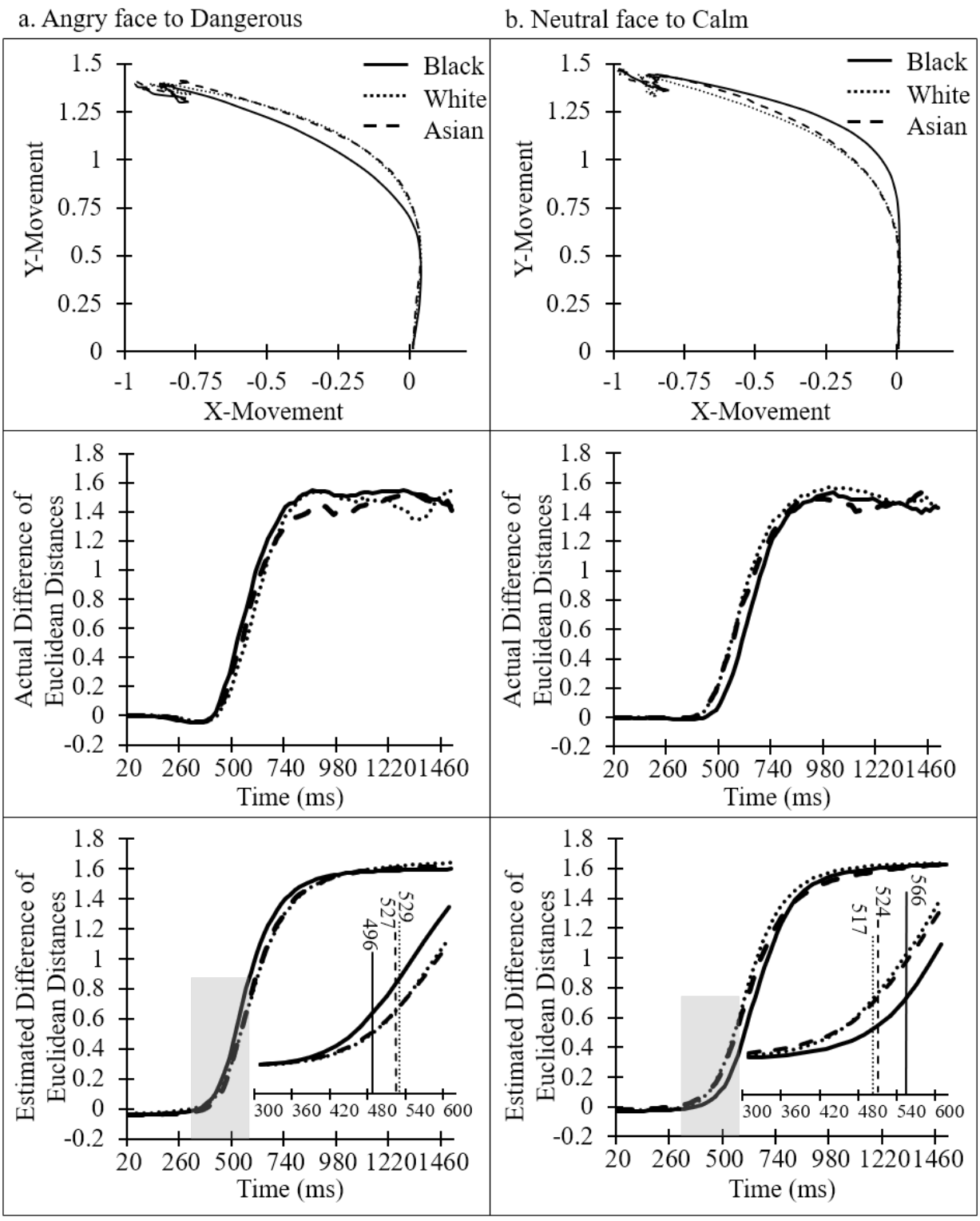

Figure 4. Average mouse trajectory (top panel), difference in Euclidean distance over time (middle panel), and estimated sigmoid from averaged Gompertz and Baranyi estimates with area of focus on TICC (bottom panel) as a function of Race for angry (left) and neutral (right) faces.

TICC in Freeman et al. (2010, Study 1).

Participants (N = 26) completed 40 trials in which they clicked a start-button to display a face which they racially categorized by moving the mouse to click “Black” or “White” (initial movement later than 400 ms yielded a message encouraging faster initial movement). The faces were Black or White and had atypical or typical race-features (e.g., morphed to have, or not, facial features of the other race). There were 40 trials with 10 trials of each stimulus-type (i.e., Race x Typicality pairing). They hypothesized that typicality would promote more efficient processing because atypical faces would activate both mental race-categories which would interfere with the judgement. We focus on their time course analysis to demonstrate what else could have been done had they had TICC. Of the 1040 total trials, they excluded 101 yielding 939 analyzed trials (the reason for the exclusions is unknown to us).

Freeman et al. (a) normed the trajectories for each participant into 101 time-steps, (b) averaged the (up to) 10 trajectories of each stimulus-type to create four average trajectories per participant (i.e., Race x Typicality), (c) calculated at each time-step the Euclidean proximity to the distractor for each average trajectory (i.e., proximity to “White” for black faces and “Black” for white faces), and (d) performed a quartile split on the time-steps and averaged within each quartile the Euclidean proximity scores for a given average trajectory, which yielded each participant’s 16 average proximity-to-distractor-scores (i.e., one for each Time-quartile x Typicality x Race paring). They submitted the scores to a 2(Typicality) x 2(Race) x 4(Time-quartile) repeated measures ANOVA. Consistent with their hypothesis was a significant Typicality x Time interaction, which indicated that typical faces evidenced less distance from the distractor in earlier than later time-quartiles than did atypical faces. A non-significant three-way interaction (p = .07) suggested the typicality effect emerged earlier for White faces and later for

Black faces.

With TICC, they would have predicted that participants would begin correctly categorizing typical faces earlier than atypical faces. Using their data, we retained the raw times and (a) averaged the (up to) 10 trajectories of each stimulus-type to create four average trajectories for each participant, (b) calculated at each time point the difference in Euclidean distance to the target and distractor for each trajectory, (c) estimated the latter sigmoid with the Gompertz and Baranyi models in SAS Proc NLIN, which converged on 103 of 104 trajectories and evidenced exceptional fit with average pseudo-R^2 of .9837 and .9829, respectively, and (d) averaged the estimates from each model yielding 4 TICCs for each participant (i.e., one for each stimulus-type). We submitted TICC scores to a 2(Typicality) x 2(Race) repeated measures ANOVA. A typicality main effect, $F(1, 24) = 28.05$, $p = .0001$, $d = 1.06$, indicated that participants began correctly categorizing typical faces earlier in time (M = 468 ms, SD = 105) than atypical faces (M = 523 ms, SD = 136). A non-significant interaction, $F(1, 24) = 3.35$, $p = .0802$, suggested the tendency to begin correctly categorizing typical faces earlier in time than atypical faces was less pronounced for White faces (480 ms vs. 515 ms), $F(1, 24) = 9.25$, $p = .0056$, $d = 0.61$, than for Black faces (457 ms vs. 531 ms), $F(1, 24) = 17.66$, $p = .0003$, $d = 0.84$. This (non-significant) interaction helps clarify Freeman et al.'s (non-significant) interaction in which the typicality effect emerged later in time for Black faces. Because participants were most apt to begin correctly categorizing typical Black faces *earliest* in time and atypical Black faces *latest* in time, the difference in Euclidian distances to the distractor (what Freeman calculated) for typical vs. atypical faces was maximized later in time for Black faces.

TICC in Hehman et al. (2014, Study 1).

Participants completed 198 trials in which they clicked a start-button to display a face

which they gender categorized by moving the mouse to click “Male” or “Female” (initial movement later than 400 ms yielded a message encouraging faster initial movement). The faces were those of 198 White politicians (80 females and 118 males). Participants subsequently rated how likely they would be to vote for each politician (1 = *Not at all* to 6 = *Very likely*). Hehman et al. hypothesized that categorization fluency would negatively impact voting, such that participants would be less likely to vote for politicians who are difficult to gender categorize. We focus on their time course analyses to demonstrate what they could have done with TICC (*N* = 29 of 33, 3 persons lacked voting data, and 1 lacked mouse data). Of the 5,742 mouse-tracking trials, they excluded 106 (1.8%) that ended in an incorrect categorization and 167 (2.9%) deemed to have an aberrant mouse trajectory, which yielded 5,469 analyzed trials.

Hehman et al. limited their time course analysis to the female politicians because shape analyses suggested that competition between the “Male” vs. “Female” label was unrelated to vote likelihood for male politicians but negatively predicted vote likelihood for female politicians. To examine the time course association with vote-likelihood they (a) calculated in 20 ms intervals the average x-coordinate of the mouse for the initial 1000 ms following face presentation, (b) computed the average vote-likelihood for each politician (i.e., averaging across all participants; personal communication Eric Hehman), and (c) yoked the latter to each participant’s x-coordinate in time for a given politician. They examined at every 20 ms time interval the correlation between the x-coordinate and average vote-likelihood (across all female politicians and participants). The correlation became (and remained) significant at the 19th interval suggesting that as early as 380 ms after face presentation horizontal movement to the correct label (female) corresponded to a greater average vote-likelihood for female politicians.

These data provide the opportunity to demonstrate the utility of TICC as a predictor in

that it enables a one-to-one pairing between a given participant's vote-likelihood for a given politician and the time that participant began correctly categorizing that politician. Keep in mind that this is not possible with the Spivey approach because each politician (i.e., stimulus) occurred once in the mouse-tracking task and the Spivey approach requires multiple trials per stimulus. Rather than limiting the analysis to female politicians, we examined all politicians. In particular, we (a) calculated at each 20 ms time point the difference in Euclidean distance to the target ("Male" for male politicians and "Female" for female politicians) and distractor ("Female" for male politicians and "Male" for female politicians) for each trajectory (i.e., trial), (b) estimated the latter sigmoid with the Gompertz and Baranyi models in SAS Proc NLIN, which converged on 5,332 (97.5%) of 5,469 trajectories and evidenced exceptional fit with average pseudo-R^2 of .9765 and .9791, respectively, and (d) averaged the estimates from each model yielding each participant a TICC for each politician.

To test whether earlier initiation of correct categorization predicts increased vote likelihood (i.e., a negative association) for male and female politicians, we conducted a multi-level regression using Proc Mixed of SAS in which we regressed vote-likelihood on politician-gender, TICC (grand-mean centered), and TICC x Politician-Gender with a random intercept for participants and Kenward-Roger degrees of freedom. A significant TICC x Politician-Gender interaction, $F(1, 5316) = 4.60$, $p = .0319$, indicated that vote likelihood was unrelated to the time of initiating correct categorization of male politicians, B = 0.000074, $t(5307) = 0.39$, $p = .6997$, but was negatively related to the time of initiating correct categorization of female politicians, B = -0.00063, $t(5312) = -2.41$, $p = .0158$. Stated otherwise, a 1 ms decrease in time to begin correctly categorizing a female politician increased the likelihood of voting for her by .00063 units (i.e., a 1 second time difference corresponds to roughly two-thirds of a vote-unit).

Of course, the latter regression effect for TICC could confound within- and between-person effects to the extent to which persons differ in TICC (i.e., overall earlier or later initiators of categorization). To distinguish within- and between-person effects (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002), we regressed vote-likelihood on politician-gender, person-centered TICC, person-mean TICC (grand-mean centered), 2-way interactions of politician-gender with each form of TICC with a random intercept and Kenwood-Rodgers degrees of freedom. Politician-gender interacted with person-centered TICC, $F(1, 5315) = 4.88$, $p = .0272$, but not person-mean TICC, $F(1, 5299) = 2.49$, $p = .1145$, indicating that the previous results were driven primarily by within-person variation in TICC. In particular, vote likelihood was again unrelated to earlier vs. later (relative to one's own mean) time of initiating correct categorization of male politicians, B = 0.000065, $t(5305) = 0.34$, $p = .7312$, but was negatively related to earlier (relative to one's own mean) time of initiating correct categorization of female politicians, B = -0.00066, $t(5310) = -2.52$, $p = .0117$. Again, a 1 ms decrease (relative to one's average) in time to begin correctly categorizing a female politician increased the likelihood of voting for her by .00066 units (i.e., a 1 second difference corresponds to two-thirds of a vote-unit).

The three data sets provided a platform for demonstrating the method for estimating TICC and the utility of TICC in testing hypotheses. In each set we estimated TICC per participant for given stimuli. In the first two sets we averaged across similar stimulus trials to estimate TICC and in the third set we estimated TICC from single trial stimulus presentations. Furthermore, we analyzed the TICC as a dependent variable in the first two sets and as a predictor variable of a stimulus yoked outcome in the third set. In the next section, we examine the applicability of this method for estimating the TICC from differently shaped mouse trajectories.

Applicability to Differently Shaped Trajectories

The Gompertz and Baranyi models estimate bacterial growth in a sigmoidal curve where growth (*y*) increases with time (*x*). We adopted those models to estimate TICC in a sigmoid that occurs in the distribution of the difference in Euclidean distance to the target versus distractor (*y*) over time (*x*). It might be assumed that application of those formulas requires an idealized mouse trajectory in which initial movement is vertical and equidistant from target and distractor and then monotonically shifts towards the target. However, such an idealized trajectory is not necessary, and the models fit many trajectories. As Figure 5 depicts, of the 6,136 trajectories for which we estimated a TICC across the data of March et al. (701 of 702 trajectory models converged), Freeman et al. (103 of 104 trajectory models converged), and Hehman et al. (5,332 of 5,469 trajectory models converged), the vast majority were exceptionally well fit by the non-linear models with a psuedo-R^2 equal to or exceeding .95. Of course, the frequency of exceptionally fitting models in Hehman et al. is inflated due to their exclusion of 2.9% of trajectories they deemed aberrant. We do not know if Freeman et al. excluded aberrant trajectories. March et al. did not exclude trials based on their trajectory and might provide a better understanding of fit. In those data, 91.5% of the modeled trajectories had a pseudo-R^2 no lower than .85 and the modal pseudo-R^2 was no lower than .95.

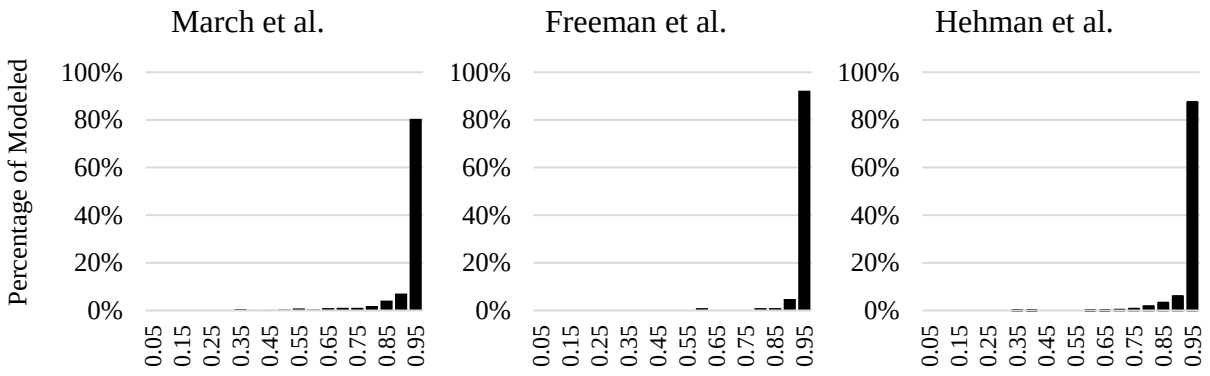

Figure 5. Histogram of Pseudo-R^2 values from March et al., Freeman et al., and Hehman et al.

To better gauge how the models fit less-than-ideal trajectories (in terms of generating a sigmoid), we explored three types of trajectories that commonly occur in the mouse tracking literature (Kieslich et al., 2020; Schoeman 2020 ; Wulff et al. 2019). One possibility (Figure 6, top panel) is for the participant to move relatively closer to the target than the distractor from movement onset onwards (i.e., resulting in a relatively straight line, or what Wulff et al. label as trajectory type 2). In this instance, the difference in Euclidean distance over time is a flatter sigmoid with a very early TICC estimate (275 ms) and a pseudo-R^2 of .99. (Keep in mind, the time axis in the figure is stimulus-locked, not response-locked.) Another possibility (Figure 6, middle panel) is that the participant might first move closer to the distractor, change their mind, and move to the target (i.e., resulting in a 7-shape, or what Wulff et al. label as trajectory type-5). In this instance, there is only one initial turn toward the correct target and TICC is estimated to occur later in time (494 ms) than in the previous pattern with a pseudo-R^2 of .90 and a noticeably lower estimate of y_{min}. A third possibility (Figure 6, bottom panel) is that the participant might move toward the target, turn to the distractor, and turn back to the target (i.e., a multi-turn trial, resulting in a messy 7-shape, not labeled by Wulff et al.). In this instance, TICC is estimated to occur even later in time (740 ms) than in the two previous patterns with a pseudo-R^2 of .75. It is also evident in the bottom panel that the estimated TICC is reflecting the initial turn toward the target and not the subsequent turn back to the target (after movement to the distractor). That subsequent turn back to the target is reflected in the actual difference in Euclidean distance of approximately zero at 1420 ms. The decreasing pseudo-R2 across the three patterns certainly reflects the increased indecision underlying the trajectory. Importantly, in these examples, the increasing estimate of TICC similarly reflects the increasing indecision with greater indecision yielding later estimates of when in time correct categorization initiates. However, an early TICC

is not necessarily diagnostic of a lack of indecision that occurs later in the trial because the TICC captures the persistent initial turn toward the target.

Whereas the non-linear models estimating TICC reflect the initial turn toward the target, a complimentary method is in development that likely captures the later turn toward the target (i.e., the "time of overt commitment" TOC; Ulbrich & Gail, unpublished manuscript). Consequently, in multi-turn trajectories it might be possible to estimate both the TICC and TOC. In such multi-turn trajectories, whereas the TICC captures the initial turn toward the target, the TOC may better reflect later indecision by capturing the last turn toward the target. In conjunction, these methods may provide unique insights into the dynamics of an often-complicated response process.

Of course, the non-linear models will not converge to yield a TICC estimate for every trajectory nor will they always provide exceptional fit to trajectories. Figure 7 provides an example of a trajectory from March et al. to which the models provided a poor fit (pseudo-R^2 = .04). Visual inspection of trajectories is sometimes recommended as a means of data exclusion (Freeman et al., 2013). Some readers might consider exclusion on the basis of poor model fit to the trajectory. We do not necessary advocate this approach, because data exclusion is a tricky issue. If data exclusion procedures are used, we recommend consideration based on multiple criteria that are established a priori (to avoid hypothesis confirmation pressures). For example, a trajectory with a low pseudo-R^2 might be retained if visual inspection of the trajectory suggests the TICC estimate captures the initial turn towards the target. Finally, we have examined the non-linear modeling method for estimating TICC in a two-choice mouse-tracking paradigm. Whether it is appropriate for other mouse-tracking paradigms, such as three or four choice paradigms, remains to be examined.

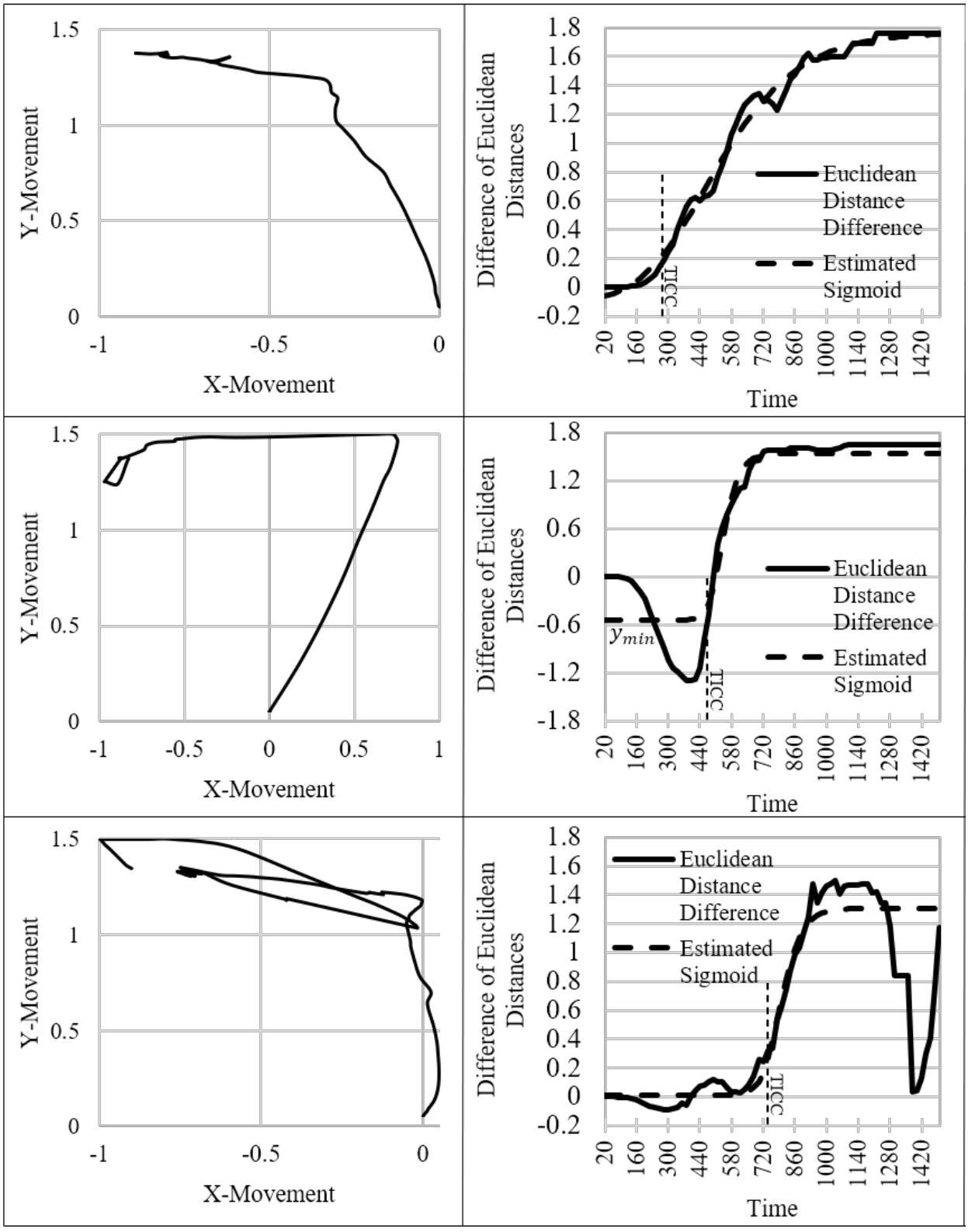

Figure 6. Examples of various X-, Y-Coordinates (left panels) and their corresponding Euclidean distance difference (solid line) and estimates sigmoid (dashed line) over time (right panels).

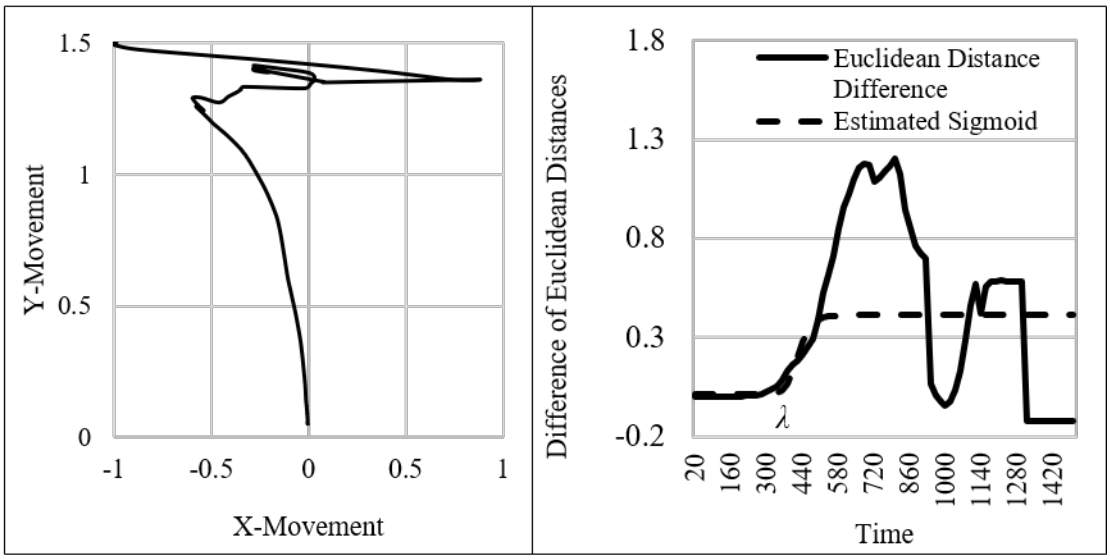

Figure 7. X-, Y-Coordinates (left panels) and their corresponding Euclidean distance difference (solid line) and estimates sigmoid (dashed line) over time (right panels) of a poor fitting model.

Closing

Mouse-tracking provides insight into mental processes that transpire across the time-course of response. Information about *when* categorization begins is relevant, but was unrealized in current mouse-tracking metrics. We adapted a bacterial growth-curve modeling approach to quantify when in time participants begin moving *relatively closer* to the chosen response (i.e., target) versus unchosen response (i.e., distractor). This time of initiating correct categorization (TICC) is an important point during the time course of categorization that begins with information search and ends with an ultimate decision.

The presented method for estimating the TICC advances previous attempts to quantify a critical time-point by providing timing information about the onset of categorization. Earlier efforts did not directly estimate time, but instead inferred a critical time via the patterning of multiple null-hypothesis tests across time-points. The TICC is an estimate of person- (or even trial-) level timing information cued to a specific stimulus and can be used as an outcome or predictor of other variables (as we demonstrated in three data sets). Given such analytic

flexibility, the TICC expands the possibilities for exploring mental processes via mouse-tracking.

Open Practices Statement

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data from March et al and analysis scripts for all presented studies are available on OSF at https://osf.io/67xzv/.

References

Baty, F. & Delignette-Muller, M. L. (2004). Estimating the bacterial lag time: Which model, which precision? *International Journal of Food Microbiology, 91*, 261-277. https://doi.org/10.1016/j.ijfoodmicro.2003.07.002

Borglin, S., Joyner, D., DeAngelis, K. M., Khudyakov, J., D'haeseleer, P., Joachimiak, M. P. & Hazen, T. (2012). Application of phenotypic microarrays to environmental microbiology. *Current Opinion in Biotechnology*, *23*, 41-48. https://doi.org/10.1016/j.copbio.2011.12.006

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition, 35*, 15–28. https://doi.org/10.3758/BF03195938

Enders, C.K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121-138. https://doi.org/10.1037/1082-989X.12.2.121

Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007). Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science, 31*, 889-909. https://doi.org/10.1080/03640210701530797

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*, 315-323. https://doi.org/10.1177/0963721417746793

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General, 137*, 673–690. https://doi.org/10.1037/a0013875

Freeman, J. B. & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods, 42,* 226-241. https://doi.org/10.3758/BRM.42.1.226

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2,* 1-6. https://doi.org/10.3389/fpsyg.2011.00059

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology, 46,* 179-185. https://doi.org/10.1016/j.jesp.2009.10.002

Freeman, J. B., Ma, Y., Han, S., & Ambady, N. (2013). Influences of culture and visual context on real-time social categorization. *Journal of Experimental Social Psychology, 49,* 206-210. https://doi.org/10.1016/j.jesp.2012.10.015

Gallivan, J. P., & Chapman, C. S. (2014). Three-dimensional reach trajectories as a probe of real-time decision-making between multiple competing targets. *Frontiers in neuroscience, 8,* 215. https://doi.org/10.3389/fnins.2014.00215

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science, 5,* 815-824. https://doi.org/10.1177/1948550614534701

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes and Intergroup Relations, 18,* 384-401. https://doi.org/10.1177/1368430214538325

Kieslich, P. J., Schoemann, M., Grage, T., Hepp, J., & Scherbaum, S. (2020). Design factors in mouse-tracking: What makes a difference?. *Behavior Research Methods*, *52*, 317-341.

https://doi.org/10.3758/s13428-019-01228-y

March, D. S., Gaertner, L., & Olson, M. A. (2021). Danger or Dislike: Distinguishing threat from valence as sources of automatic anti-Black bias. *Manuscript Under Review.*

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*, 277-294. https://doi.org/10.1146/annurev-psych-010814-015148

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage.

Scherbaum, S., & Dshemuchadse, M. (2020). Psychometrics of the continuous mind: Measuring cognitive sub-processes via mouse tracking. *Memory & Cognition*, *48*, 436-454. https://doi.org/10.3758/s13421-019-00981-x

Schoemann, M., Lüken, M., Grage, T., Kieslich, P. J., & Scherbaum, S. (2019). Validating mouse-tracking: How design factors influence action dynamics in intertemporal decision making. *Behavior Research Methods, 51*, 2356-2377. https://doi.org/10.3758/s13428-018-1179-4

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2020) Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. Psychonic Bulletin & Review. https://doi.org/10.3758/s13423-020-01851-3

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Science, 102,* 10393-10398. https://doi.org/10.1073/pnas.0503903102

Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to

the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*, 122-134. https://doi.org/10.1177/0956797614559543

Ulbrich, P., & Gail, A. (unpublished manuscript). The Cone Method: Inferring Decision Times from Single-Trial 3D Movement Trajectories in Choice Behavior. *BioRxiv*. https://doi.org/10.1101/2020.08.01.232314

Wulff, D. U., Haslbeck, J. M. B., Kieslich, P. J., Henninger, F., & Schulte-Mecklenbeck, M. (2019). Mousetracking: Detecting types in movement trajectories. In M. Schulte-Mecklenbeck, A. Kühberger, & J. J. Johnson (Eds.), A handbook of process tracing methods (2nd ed., pp. 131-145). Routledge. https://doi.org/10.4324/9781315160559-10