METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Filling the Gap Between Implicit Associations and Behavior: A Linear Mixed-Effects Rasch Analysis of the Implicit Association Test

Ottavia M. Epifania [1], Pasquale Anselmi [1], Egidio Robusto [1]

[1] *Department of Philosophy, Sociology, Education, and Applied Psychology, University of Padova, Padova, Italy.*

Methodology, 2022, Vol. 18(3), 185–202, https://doi.org/10.5964/meth.7155

Received: 2021-07-15 • **Accepted:** 2022-09-01 • **Published (VoR):** 2022-09-30

Handling Editor: Jost Reinecke, Bielefeld University, Bielefeld, Germany

Corresponding Author: Ottavia M. Epifania, Department of Philosophy, Sociology, Education, and Applied Psychology, Via Venezia 14, Padova, Italy. E-mail: ottavia.epifania@unipd.it

Supplementary Materials: Data [see Index of Supplementary Materials]

Abstract

The measure obtained from the Implicit Association Test (IAT; Greenwald et al., 1998. DOI: 10.1037/0022-3514.74.6.1464) is often used to predict people's behaviors. However, it has shown poor predictive ability potentially because of its typical scoring method (the D score), which is affected by the across-trial variability in the IAT data and might provide biased estimates of the construct. Linear Mixed-Effects Models (LMMs) can address this issue while providing a Rasch-like parametrization of accuracy and time responses. In this study, the predictive abilities of D scores and LMM estimates were compared. The LMMs estimates showed better predictive ability than the D score, and allowed for in-depth analyses at the stimulus level that helped in reducing the across-trial variability. Implications of the results and limitations of the study are discussed.

Keywords

implicit association test, Rasch model, log-normal model, mixed-effects models, attitude-behavior gap

The Implicit Association Test (IAT; Greenwald et al., 1998) is one of the most used measures for the implicit assessment of socio-psychological constructs. The main fields of application are in social psychology, where the IAT is often employed to indirectly investigate the attitudes towards different social groups. Additionally, the IAT is used to assess food and brand preferences (see Epifania et al., 2022a, for a review of the main fields of application of the IAT). In both fields, the measure provided by the IAT is

used to predict behavioral outcomes, such as intergroup relations (e.g., Dovidio et al., 2002) or food choice (e.g., Perugini, 2005). However, the IAT has shown poor ability to predict behavioral outcomes (e.g., Meissner et al., 2019), potentially because of its typical scoring method (i.e., the so-called *D* score; Greenwald et al., 2003). If the poor ability of the IAT to predict behaviors is ascribable to its typical scoring method, the estimates obtained with more statistically sound approaches should result in better predictions. In this contribution, a Rasch analysis based on Linear Mixed-Effects Models (LMMs) is introduced to address the across-trial variability in the IAT data and to obtain reliable measures for accurate predictions of behaviors.

The IAT assesses the strength of the associations between targets and evaluative dimensions by considering the speed and accuracy with which prototypical exemplars of two targets (e.g., *Coke* and *Pepsi* images in a Coke-Pepsi IAT) and two evaluative dimensions (*Good* and *Bad* attributes) are assigned to their own category in two contrasting conditions. In one condition, *Coke* and *Good* exemplars are assigned with the same key, while *Pepsi* and *Bad* exemplars are assigned with the opposite key. In the contrasting condition, *Pepsi* and *Good* exemplars are assigned with the same key, while *Coke* and *Bad* exemplars are assigned with the opposite key. The task is expected to be easier (i.e., responses should be faster and more accurate) in the condition consistent with one's own automatically activated association. The *D* score (Greenwald et al., 2003) is usually employed to express the IAT effect (i.e., the difference in the performance between the two conditions). It is an effect size measure obtained by standardizing the difference between the average response time in the two conditions by the standard deviation computed on the pooled trials of both conditions.

The IAT effect as expressed by the *D* score has been found to have poor ability to predict behaviors. This can be ascribed to different factors, including the measure provided by the *D* score, the construct assessed by the IAT (Meissner et al., 2019), and the type of behavioral outcomes (Perugini, 2005). Additionally, the fully-crossed structure of the IAT (Westfall et al., 2014) might compromise the predictive ability of its measure. If the fully-crossed design of the IAT and its related sources of dependency are not properly addressed, biased estimates are obtained, the importance of experimental effects is confused with random noise, and the probability of committing Type I error is inflated (Judd et al., 2017; Wolsiefer et al., 2017). Averaging across trials in each associative condition, the *D* score is highly sensitive to the across-trial variability related to stimuli heterogeneity, and it cannot address the fully-crossed design of the IAT (Wolsiefer et al., 2017). This can be accounted for by employing Linear Mixed Effect-Models (LMMs) with appropriate random structures. Additionally, LMMs allow for obtaining parametrizations from accuracy and log-time responses that are conceptually close to the Rasch (Rasch, 1960) and the log-normal (van der Linden, 2006) models, respectively. These models disentangle the unique contribution of the respondent and the stimulus to the observed response, hence providing fine-grained information at both levels.

Information at the stimulus level allows for investigating the contribution of each stimulus to the IAT effect as well as the representativeness of each stimulus. Indeed, stimulus representativeness of its own category is a key feature for a correct functioning of the IAT (Bluemke & Friese, 2006; Nosek et al., 2005). Selecting the most informative and representative stimuli can help in reducing the across-trial variability, and could allow for designing better functioning and briefer IATs.

In this study, the predictive abilities of the estimates obtained with LMMs and the *D* score are compared. The predictive abilities of *D* scores computed on all stimuli and *D* scores computed only on the most (or the least) informative stimuli are compared, as well. To these ends, an IAT for the implicit assessment of the chocolate preference was used (Chocolate IAT). The most and the least informative stimuli are identified by considering the difference in their parameters between conditions (see e.g., Anselmi et al., 2013). Stimuli showing a higher difference in their parameters between conditions are considered to be more informative than those with a smaller difference in their parameters between conditions.

Method

Participants

Seventy-six university students ($F = 71.05\%$, Mean age = 24.02 ± 2.88 years) volunteered to take part in the study. Respondents did not receive any incentives for their participation.

Materials and Procedure

The script used for running the experiment, the stimuli, and the data can be accessed at the Supplementary Materials section. Twenty-six attribute stimuli (13 *Good* and 13 *Bad* exemplars) and 7 chocolate images graphically modified to represent either dark or milk chocolate (7 *Dark* and 7 *Milk* chocolate images) were used. Sixty trials were presented in each associative condition (i.e., Dark-Good/Milk-Bad–DGMB–and Milk-Good/Dark-Bad–MGDB–conditions). No feedback followed incorrect responses.

The chocolate preferences were explicitly investigated with two items (i.e., *How much do you like dark chocolate?* and *How much do you like milk chocolate?*) evaluated on a 6-point Likert-type scale (0—*Not at all*, 5—*Very much*). Respondents were asked about their food habits and behaviors through 6 items (example item: *I am usually on a diet*, Cronbach's $\alpha = 0.80$) rated on a 4-point agreement Likert-type scale (1—*Strongly disagree*, 4—*Strongly agree*). High scores indicate high care for food habits. At the end of the experiment, participants were offered with dark or milk chocolate. Their choices were registered after they left the laboratory.

Data Cleaning and D Score

Exclusion criteria based on accuracy (Nosek et al., 2002) and time responses (Greenwald et al., 2003) were applied. The IAT was scored with the *D*4 algorithm (Greenwald et al., 2003), which was computed with the online app DscoreApp (Epifania et al., 2020). Positive *D* scores denote a preference for dark chocolate relative to milk chocolate.

Model Specifications

According to the Rasch model (Rasch, 1960), the observed accuracy response of respondent p ($p \in \{1, \ldots, P\}$) to stimulus s ($s \in \{1, \ldots, S\}$) depends on respondent's ability (i.e., the respondent's ability parameter θ) and stimulus difficulty (i.e., the stimulus difficulty parameter b). In the IAT, the higher the ability parameter θ of respondent p, the higher the ability of respondent p to perform the categorization task. The higher the difficulty parameter b of stimulus s, the lower the probability of s to be assigned to the correct category. The probability of a correct response of respondent p to stimulus s depends on the distance between respondent and stimulus parameters (i.e., $\theta_p - b_s$). It is larger than .50 when $\theta_p > b_s$, smaller than .50 when $\theta_p < b_s$, and equal to .50 when $\theta_p = b_s$.

Similar to the Rasch model, in the log-normal model (van der Linden, 2006) the observed log-time response depends on the characteristics of the respondent (speed parameter τ) and those of the stimulus (time intensity parameter δ). In the IAT case, the lower the speed parameter τ of respondent p, the higher the time spent by respondent p on the task (i.e., lower speed). The lower the time intensity parameter δ of stimulus s, the lower the time respondents spend in responding to stimulus s. The expected log-time response is a function of the distance between respondent and stimulus parameters (i.e., $\delta_s - \tau_p$). The expected log-time response is lower than, faster than, and equal to the observed log-time response when $\delta_s > \tau_p$, $\delta_s < \tau_p$, and $\delta_s = \tau_p$, respectively.

Rasch-like and log-normal parametrizations can be obtained by using Generalized Linear Mixed-Effects Models (GLMMs) with *logit* link functions applied to accuracy responses and Linear Mixed Effects Models (LMMs) applied to log-time responses, respectively. In these applications, respondent and stimulus parameters are summed (i.e., $\theta_p + b_s$ and $\delta_s + \tau_p$). This parametrization of the accuracy responses is consistent with that of linear test models (LLTM, see e.g., Fischer, 1973; Scheiblechner, 1972). The higher the value of stimulus parameter b, the easier stimulus s is (i.e., the higher the number of correct responses registered on stimulus s is), such that parameter b is considered as an easiness parameter. The lower the value of parameter τ, the faster respondent p is. The suitability and usefulness of this approach for analyzing IAT data has already been proved (e.g., Epifania et al., 2022b).

Rasch-like and log-normal parametrizations depend on the factors specified as random, which account for the variability in the data. The fixed intercept is set at 0 (i.e., none of the levels of the fixed slope—the associative condition—is taken as the reference

level). Further details on the procedure and on the random structures of the models are reported in the Appendix. Table 1 summarizes the Rasch-like and log-normal parameters attainable from each model random structure.

Table 1

Rasch-Like and Log-Normal Parametrizations

	Rasch-Like Parametrization		Log-Normal Parametrization	
Model	**Respondents**	**Stimuli**	**Respondents**	**Stimuli**
1	Overall (θ_p)	Overall (b_s)	Overall (τ_p)	Overall (δ_s)
2	Overall (θ_p)	Condition–specific (b_{sc})	Overall (τ_p)	Condition–specific (δ_{sc})
3	Condition–specific (θ_{pc})	Overall (b_s)	Condition–specific (τ_{pc})	Overall (δ_s)

Note. $p \in \{1, \ldots, P\}$, $s \in \{1, \ldots, S\}$, $c \in \{1, \ldots, C\}$ denote any respondent, stimulus, condition (P, S, and C are the number of respondents, stimuli, and conditions, respectively.)

In Model 1, the random intercepts of respondents and stimuli are specified to account for the between—respondents and the between—stimuli variabilities across–conditions. This model yields overall respondent (θ_p or τ_p) and stimulus (b_s or δ_s) parameters across associative conditions. Model 1 is expected to be the best fitting one when low between–conditions variability is observed at both respondent and stimulus levels (i.e., neither respondents' performance nor stimuli functioning vary between associative conditions).

Specifying stimulus random slopes in associative conditions and respondent's random intercepts across conditions, Model 2 accounts for the within–stimuli between–conditions variability and the between–respondents across–conditions variability. This model yields overall respondent (θ_p or τ_p) and condition–specific stimulus (b_{sc} or δ_{sc}, where c denotes the associative condition) parameters. Model 2 is expected to be the best fitting model when high within–stimuli between–conditions variability is observed. This suggests that the IAT effect is mostly due to variations in stimuli functioning between conditions. The difference between condition–specific stimulus estimates allows for investigating the contribution of each stimulus to the IAT effect.

Model 3 addresses the within–respondents between–conditions variability and the between– stimuli across–conditions variability by specifying respondent's random slopes in associative conditions and stimulus random intercepts across conditions. Model 3 yields condition–specific respondent (θ_{pc} or τ_{pc}) and overall stimulus (b_s or δ_s) parameters. Model 3 is expected to be the best fitting model when high within–respondents between–conditions variability is observed, this suggesting that the IAT effect is mostly due to the changes in respondents' performance between conditions. The difference between respondent condition–specific estimates allows for investigating the bias on respondents' performance due to the IAT associative conditions.

The models were applied to the Chocolate IAT data. In what follows, the models applied to accuracy responses are identified by a capital "A". Those applied to log-time responses are identified by a capital "T". No correction was applied on the incorrect time responses for estimating the log-normal models. Models were fitted with the lme4 package (Bates et al., 2015b) in R (Version 3.5.1, R Core Team, 2018). Simple R scripts for estimating these models from any IAT are available at the Supplementary Materials section.

Results

Two participants showed more than 25% of incorrect responses in at least one associative condition (Nosek et al., 2002). The final sample consisted of 74 participants (F = 71.62%, Mean age = 24.08 ± 2.88 years). The 41.90% of the participants chose milk chocolate.

Accuracy Models

Model comparison is reported in the top panel of Table 2. BIC suggests a better fit of Model A1 compared to model A2, whereas AIC, Log-likelihood, and Deviance suggest a better fit of Model A2. Thus, Model A2 was chosen. This model provides overall Rasch-like respondent ability (θ_p) and condition–specific stimulus easiness (b_{MGDB} and b_{DGMB}) estimates. In this application, the ability estimates θ_p can be considered as accuracy-based measures of the respondents' preference. Condition MGDB showed higher probability of correct responses (*log-odds* = 3.67, SE = 0.14, z = 26.15, p < .001) than condition DGMB (*log-odds* = 2.61, SE = 0.10, z = 27.26, p < .001). Between–respondents variability was 0.33. Stimuli showed higher variability in the MGDB condition (σ^2 = 0.21) than in the DGMB condition (σ^2 = 0.01). The condition–specific stimulus random effects were weakly correlated (r = .20).

Table 2

Model Comparison Between Accuracy (Top Panel) and Log-Time (Bottom Panel) Models

Model	AIC	BIC	Log-Likelihood	Deviance
Accuracy				
A1	3627.70	3656.10	−1809.90	3619.70
A2	3625.58	3668.10	−1806.80	3613.60
A3		Failed to converge		
Log-time				
T1	7856.45	7891.91	−3923.23	7846.45
T2		Aberrant estimates		
T3	7159.23	7208.87	−3572.62	7145.23

The condition–specific easiness estimates are reported in Table 3.

Table 3

Condition–Specific Easiness Estimates (b_{sc}) and Overall Time Intensity Estimates (δ_s) of the Stimuli

Good **attributes**	b_{DGMB}	b_{MGDB}	$b_{DGMB}-b_{MGDB}$	δ_s	***Bad*** **attributes**	b_{DGMB}	b_{MGDB}	$b_{DGMB}-b_{MGDB}$	δ_s
joy[a]	2.62	4.02	−1.40	0.01	hate[a]	2.59	3.85	−1.26	0.01
happiness[a]	2.64	4.03	−1.39	0.02	failure[a]	2.68	3.93	−1.25	0.07
pleasure[a]	2.56	3.70	−1.15	0.01	terrible[a]	2.64	3.89	−1.24	0.04
peace	2.64	3.77	−1.14	−0.03	disaster	2.66	3.90	−1.24	0.07
heaven	2.63	3.77	−1.14	0.08	bad	2.58	3.73	−1.15	0.07
marvelous	2.66	3.79	−1.13	0.05	horrible	2.62	3.76	−1.14	0.05
laughter	2.67	3.76	−1.10	0.06	evil	2.63	3.74	−1.11	0.10
good	2.66	3.74	−1.08	0.01	disgust	2.60	3.70	−1.11	0.01
glory	2.57	3.57	−1.00	0.02	nasty	2.59	3.33	−0.74	0.04
love	2.62	3.58	−0.96	0.02	ugly	2.60	3.32	−0.72	−0.01
excellent[b]	2.64	3.59	−0.95	0.01	pain[b]	2.58	3.23	−0.65	0.05
beauty[b]	2.61	3.46	−0.85	0.02	annoying[b]	2.58	3.05	−0.47	0.08
wonderful[b]	2.62	3.45	−0.83	0.09	agony[b]	2.57	2.49	0.08	0.04
M (*SD*)	2.63 (0.03)	3.71 (0.17)	−1.09 (0.17)	0.03 (0.03)	*M* (*SD*)	2.61 (0.03)	3.53 (0.41)	−0.92 (0.40)	0.05 (0.03)
Dark **Chocolate**	b_{DGMB}	b_{MGDB}	$b_{DGMB}-b_{MGDB}$	δ_s	***Milk*** **Chocolate**	b_{DGMB}	b_{MGDB}	$b_{DGMB}-b_{MGDB}$	δ_s
Dark5[a]	2.56	3.94	−1.38	−0.12	Milk3[a]	2.60	3.95	−1.35	−0.04
Dark2[a]	2.60	3.82	−1.23	−0.11	Milk6[a]	2.66	3.99	−1.33	−0.04
Dark6[a]	2.55	3.72	−1.16	−0.10	Milk4[a]	2.53	3.80	−1.27	−0.04
Dark4	2.62	3.62	−1.00	−0.07	Milk2	2.57	3.61	−1.04	−0.06
Dark3[b]	2.58	3.53	−0.95	−0.08	Milk5[b]	2.62	3.64	−1.02	−0.05
Dark7[b]	2.58	3.41	−0.83	−0.07	Milk1[b]	2.62	3.62	−1.01	−0.03
Dark1[b]	2.49	3.27	−0.78	−0.11	Milk7[b]	2.54	3.49	−0.95	−0.04
M (*SD*)	2.57 (0.03)	3.62 (0.22)	−1.05 (0.20)	−0.10 (0.02)	*M* (*SD*)	2.59 (0.05)	3.73 (0.17)	−1.14 (0.17)	−0.04 (0.01)

Note. DGMB: Dark-Good/Milk-Bad condition; MGDB: Milk-Good/Dark-Bad condition. Rows are ordered by increasing values of $b_{DGMB}-b_{MGDB}$. The units of the easiness estimates are the log-odds, the units of the time intensity estimates are the log-seconds.

[a]Stimuli that, according to the condition–specific easiness estimates, contributed the most to the IAT effect.

[b]Stimuli that, according to the condition–specific easiness estimates, contributed the least to the IAT effect.

Stimuli were easier in the MGDB condition than in the DGMB one, $M_{MGDB} = 3.64 \pm 0.29$, $M_{DGMB} = 2.60 \pm 0.04$; $t(40) = -21.97$, $p < .001$, 95% CI [−1.13, −0.94]. A linear model was specified to investigate the effect of the stimulus categories on the difference between condition–specific easiness estimates, which can be considered as an accuracy-based measure of the IAT effect. An overall significant effect of the stimulus categories was found, $F(4, 36) = 139.80$, $p < .001$, *Adjusted* $R^2 = 0.93$. *Milk* and *Good* exemplars contributed

the most to the IAT effect, $B_{Milk} = -1.13$, $SE = 0.11$, $t(36) = -10.84$, $p < .001$; $B_{Good} = -1.09$, $SE = 0.08$, $t(36) = -14.10$, $p < .001$). *Bad* and *Dark* exemplars contributed the least ($B_{Bad} = -0.92$, $SE = 0.07$, $t(36) = -11.98$, $p < .001$; $B_{Dark} = -1.05$, $SE = 0.11$, $t(36) = -9.97$, $p < .001$.

Log-Time Models

Model comparison is reported in the bottom panel of Table 2. Model T3 was chosen, providing overall stimulus time intensity (δ_s) and respondent condition–specific speed estimates (τ_{MGDB} and τ_{DGMB}) of the log-normal model. Responses were faster in the MGDB condition ($B = -0.36$, $SE = 0.02$, $t = -15.01$) than in the DGMB condition ($B = -0.12$, $SE = 0.03$, $t = -4.28$). The between–stimuli variability was extremely low ($\sigma^2 = 0.004$). Respondents showed similar variabilities in DGMB and MGDB conditions ($\sigma^2_{DGMB} = 0.05$; $\sigma^2_{MGDB} = 0.03$), and their random effects were moderately correlated ($r = .40$). A linear model was specified to investigate the effect of the stimulus categories on the time intensity estimates (Table 3). An overall significant effect of the stimulus categories was found, $F(4, 36) = 37.41$, $p < .001$, *Adjusted* $R^2 = 0.78$. The exemplars of both targets required the least amount of time to get a response ($B_{Dark} = -0.09$, $SE = 0.01$, $t(36) = -8.99$, $p < .001$; $B_{Milk} = -0.04$, $SE = 0.01$, $t(36) = -4.09$, $p < .001$), whereas exemplars of both evaluative dimensions required the largest amount of time ($B_{Bad} = 0.05$, $SE = 0.01$, $t(36) = 6.20$, $p < .001$; $B_{Good} = 0.03$, $SE = 0.01$, $t(36) = 3.70$, $p < .001$).

Relationship Between Model Estimates, D Scores, and Explicit Measures

A speed-differential was obtained by taking the difference between the condition–specific speed estimates, which can be considered as a latency-based measure of the IAT effect. Positive values indicated higher speed in the DGMB condition than in the MGDB condition. Results of Pearson's correlations between explicit measures, *D* scores, and model estimates are reported in Table 4. Explicit chocolate evaluations strongly correlated with *D* scores and condition–specific speed estimates.

The accuracy-based measure of the respondent's preference correlated neither with explicit chocolate evaluations nor with any of the condition–specific speed estimates or the *D* score. As such, it appears these estimates cannot be considered as an indicator of the implicit preference of the respondents. High speed in the MGDB condition correlated with positive milk chocolate evaluations, and not with the dark chocolate evaluations. Similarly, high speed in the DGMB condition correlated with positive dark chocolate evaluations, and not with the milk chocolate evaluations. This suggests that the performance in each associative condition is mostly driven by the associations between one of the two chocolates and positive attributes. In this sense, the like for each of the two chocolates has a major importance in influencing the responses.

Table 4

Correlation Between Model Estimates, Explicit Measures, and D Score

Condition	1	2	3	4	5	6
1 - Explicit Milk						
2 - Explicit Dark	−0.51***					
3 - *D* score	−0.43***	0.51***				
4 - τ_{DGMB}	0.12	−0.43***	−0.60***			
5 - τ_{MGDB}	−0.36**	0.14	0.42***	0.42***		
6 - θ_p	0.01	0.18	0.06	0.07	0.18	
7 - *Speed-differential*	−0.41***	0.55***	0.95***	−0.67***	0.39***	0.07

Note. τ: speed estimate; θ: Accuracy-based measure of respondents' preference, DGMB: Dark-Good/Milk-Bad condition; MGDB: Milk-Good/Dark- Bad condition; *Speed-differential*: $\tau_{MGDB} - \tau_{DGMB}$.
**$p < .01$. ***$p < .001$.

Choice Prediction

The predictive abilities of model estimates and *D* scores were compared. Two data sets were created from the full-length data set by selecting the responses to the three stimuli of each category that contributed the most (stimuli in Table 3 marked with [a]) or the least (stimuli in Table 3 marked with [b]) to the IAT effect. The *D4* algorithm was computed on both data sets. The predictive abilities of differential measures (i.e., *D* scores and speed-differential) and of their single components (i.e., M_{MGDB} and M_{DGMB} of the *D* scores, τ_{DGMB} and τ_{MGDB} of the speed-differential) were investigated. All predictors were checked for collinearity by computing Variance Inflation Factors (*VIF*s). The *D* score was collinear with the speed differential, the two condition–specific speed estimates, and the condition–specific average response times (*VIF*s > 10). Condition–specific speed estimates were not collinear between each other (*VIF*s < 4.00), but they were collinear with condition–specific average response times. Condition–specific speed and average response times, *D* score, and speed differential were not collinear with food habits and preference estimates (*VIF*s < 4.00). Given the high collinearity between the predictors (i.e., the *D* score and the other time-based predictors, namely the condition–specific speed estimates, the condition–specific average response times, and the speed differential), they were entered in separate models. As such, eight logistic regression models were specified. Preference estimates and food habits of the respondents were included in all starting models. Either the *D* score, the speed differential, the condition–specific speed estimates, or the condition–specific average response times were included in the same model. Relevant predictors were selected with backward deletion. Model general accuracy (i.e., percentage of choices correctly identified by the model), model dark chocolate choice

(DCC) accuracy (i.e., percentage of DCCs correctly identified by the model), and model milk chocolate choice (MCC) accuracy (i.e., percentage of MCCs correctly identified by the model) were computed on the models resulting from backward deletion (Table 5).

Table 5

Choice Prediction: Models Resulting After Backward Deletion

Predictors	*B*	*SE*	*Nagelkerke* R^2	*General*	*DCC*	*MCC*
Intercept	−1.65**	0.51	0.26	66%	70%	61%
D score	−2.03***	0.60				
Intercept	−1.65***	0.48	0.26	68%	72%	61%
Speed-differential	−5.02***	1.43				
Intercept	−1.76***	0.52	0.30	70%	74%	65%
D score (Best)	−2.07***	0.58				
Intercept	−1.23***	0.42	0.18	69%	72%	65%
D score (Worst)	−1.40***	0.47				
			Single components			
Intercept	−0.23	1.36	0.27	65%	74%	52%
M_{DGMB}	0.00**	0.01				
M_{MGDB}	−0.01**	0.01				
Intercept	−2.05*	0.74	0.27	72%	74%	68%
τ_{DGMB}	4.73***	1.48				
τ_{MGDB}	−5.99***	1.98				
Intercept	−0.17	1.61	0.30	65%	74%	52%
M_{DGMB} (Best)	0.00***	0.01				
M_{MGDB} (Best)	−0.01*	0.01				
Intercept	0.61	1.23	0.16	64%	77%	45%
M_{DGMB} (Worst)	0.00*	0.01				
M_{MGDB} (Worst)	0.00*	0.01				

Note. Best: Highly contributing stimuli data set; Worst: Lowly contributing stimuli data set; τ: Speed; *Speed-differential*: τ_{MGDB}–τ_{DGMB}; DGMB: Dark-Good/Milk-Bad condition; MGDB: Milk-Good/Dark-Bad condition; *General*: General accuracy of chocolate choice predictions; *DCC*: Dark Chocolate Choice Accuracy; *MCC*: Milk Chocolate Choice Accuracy.
*$p < .05$. **$p < .01$. ***$p < .001$.

Speed-differentials and *D* scores resulted in similar predictive accuracies. "Best" and "Worst" data sets *D* scores provided more accurate predictions than full data set *D* scores. The "Best" data set *D* scores explained the highest proportion of variance. Condition–specific speed estimates resulted in the highest MCC accuracy.

Final Remarks

This study investigated whether the predictive ability of the IAT could be enhanced with statistical models able to account for its fully-crossed structure. The results suggested that the proposed modeling framework can improve the predictive ability of the IAT while providing information on respondent's performance and stimulus functioning. This information can be further employed to reduce the across-trial variability due to stimuli heterogeneity, thus leading to better functioning, more informative, and potentially briefer IATs.

The stimulus functioning in respect to both its own category and other categories can be investigated through stimulus time intensity estimates. The within–category variability allows for identifying the most and the least representative stimuli of each category, whereas the between–category variability suggests different times for processing target and attribute exemplars that potentially contribute to the across-trial variability.

Condition–specific easiness estimates suggested that the IAT effect in the Chocolate IAT was mostly driven by *Good* and *Milk* exemplars. Consistently, the correlations between condition–specific speed estimates and differential measures pointed at a major influence of the speed in the MGDB condition. The correlations between speed estimates and explicit chocolate evaluations further suggested that the performance in each condition was mostly influenced by positive attributes. As such, it can be speculated that the IAT effect is mostly driven by a milk chocolate preference, but the performance in each condition is mostly influenced by the associations of positive attributes with one of the two chocolates. The ability of the model estimates to disentangle the component(s) mostly involved in the performance at the IAT might have a high resonance in both marketing and applied social psychology. In the former field, it can clarify whether the obtained results are mostly due to the preference for one of two contrasting brands and help in designing *ad hoc* marketing campaigns. In the latter one, it can disentangle whether the performance at the IAT is mostly due to in-group preference rather than outgroup derogation. Understanding whether individuals more easily associate the in-group with positive attributes rather than the outgroup with negative ones has important practical implications.

Previous studies have stressed the sensitivity of the IAT to the stimulus properties, suggesting that valid IATs can be obtained with a small number of highly informative and representative stimuli (Bluemke & Friese, 2006; Nosek et al., 2005). In this application, the selection of highly contributing stimuli allowed for reducing the across-trial variability, such that the number of trials was minimized while the information that could be gathered from the IAT was maximized. This unveils the possibility of reducing the length of the IAT without losing information and/or impairing its validity. Reducing the stimuli heterogeneity also resulted in *D* scores better able to predict the behavioral outcome. The *D* scores computed on the most informative data set explained the highest proportion of variance and provided better predictions than the *D* scores computed on

the full-length data set. Interestingly, also the *D* scores computed on the least informative data set better predicted the choice than the full-length *D* scores. We speculate that by reducing the stimuli heterogeneity and the across trial variability, more reliable *D* scores can be obtained because the sources of error variance are accounted for. Being more reliable, the *D* scores obtained on reduced data sets can better predict behavioral outcomes than those obtained on full data sets, which are affected by error variance. This result might further stress the sensitivity of the *D* score to the across-trial variability. However, further investigations on this topic are needed.

In this study, the target categories (i.e., dark chocolate and milk chocolate) were quite homogeneous. The modeling framework helped in highlighting the stimuli with a different functioning in respect to the stimuli belonging to the same category and those that mostly contributed to the IAT effect (i.e., the stimuli that presented a high difference in their easiness estimates between conditions). This information contributed to get a better understanding of the IAT measure, and to reduce the across-trial variability, leading to a better prediction of the behavioral outcome. When target categories are more heterogeneous (as it could be, e.g., race), the proposed modeling framework can identify the malfunctioning stimuli and those that mostly contribute to the IAT effect (Epifania et al., 2021). A reduction of the across-trial variability can be expected also in the case of heterogeneous categories, but it might not directly result in better predictions of behavioral outcomes. In these cases, the heterogeneity of the categories might require a larger collection of stimuli to appropriately represent them and to efficiently predict behavioral outcomes of interest. Future studies should investigate the functioning of the proposed modeling framework when heterogeneous categories are used.

The comparisons between the full-length IAT and the short IATs based on the responses from the same starting data set constitutes the main limitation of the study. In future studies, two IATs could be designed, one including only highly representative stimuli, the other one including only poorly representative stimuli. If the results are replicated with these IATs, further evidence on the importance of the representativeness of the stimuli and about the *D* score sensitivity to the across-trial variability would be obtained.

Other models that can concurrently account for accuracy and time responses have been applied to the IAT data, namely the Diffusion Model (DM; Klauer et al., 2007) and the Discrimination Association Model (DAM; Stefanutti et al., 2013, see also the four-counter DAM; Stefanutti et al., 2020). DM and DAM consider the performance of the respondents at the IAT as the result of different processes, each of which is expressed by its own parameter. As such, both models provide in-depth information concerning the individual differences of the respondents. However, no information at the single stimulus level is available, but only at the stimulus categories level. On the other hand, the modeling framework introduced in this contribution results in fine-grained information also at the individual stimulus level, which in turn allows for the investigation of the stimuli representativeness of their own category as well as of their contribution to the

IAT effect. A limitation of this study is that it does not provide a direct comparison between the information resulting from the DAM or the DM and that resulting from the modeling framework proposed here. Such a comparison could be of interest for future studies.

The convergence failure of Model A3 and the aberrant estimates obtained with Model T2 raise concerns and should be considered as a potential drawback of the modeling framework introduced in this contribution. Convergence failure or aberrant estimates suggest that the model could not find a solution, usually because of a lack of variability in the data (i.e., the random structure of the model requires a higher variability than that observed in the data, Bates et al., 2015a). The poor variability in the accuracy performance of the respondents (SD = 0.11) might have caused the convergence failure of Model A3. Similarly, the poor variability in the response times of the stimuli (SD = 0.02) might have caused the degenerate solution of Model T2.

Funding: The research leading to these results has been carried out in the frame of the project "New frontiers in test development" under the agreement code "C59C20000050005".

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that they have no conflicts of interest to disclose.

Data Availability: Data is freely available at Supplementary Materials.

Supplementary Materials

The supplementary materials provided include the R scripts used for running the experiment, the stimuli, and the data (see Index of Supplementary Materials below).

Index of Supplementary Materials

Epifania, O. M., Anselmi, P., & Robusto, E. (2022). *Supplementary materials to "Filling the gap between implicit associations and behavior: A linear mixed-effects Rasch analysis of the Implicit Association Test"* [R scripts, stimuli, data]. OSF. https://osf.io/54qat/

References

Anselmi, P., Vianello, M., Voci, A., & Robusto, E. (2013). Implicit sexual attitude of heterosexual, gay and bisexual individuals: Disentangling the contribution of specific associations to the overall measure. *PLoS One, 8*(11), Article e78990. https://doi.org/10.1371/journal.pone.0078990

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015a). *Parsimonious mixed models.* arXiv. https://doi.org/10.48550/arXiv.1506.04967

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology, 42*(2), 163–176. https://doi.org/10.1016/j.jesp.2005.03.004

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*(2), 1–18. https://doi.org/10.18637/jss.v020.i02

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62–68. https://doi.org/10.1037/0022-3514.82.1.62

Epifania, O. M., Anselmi, P., & Robusto, E. (2020). DscoreApp: A shiny web application for the computation of the Implicit Association Test D-score. *Frontiers in Psychology, 10*, Article e2938. https://doi.org/10.3389/fpsyg.2019.02938

Epifania, O. M., Robusto, E., & Anselmi, P. (2021). Rasch gone mixed: A mixed model approach to the Implicit Association Test. *TPM. Testing, Psychometrics, Methodology in Applied Psychology, 28*(4), 467–483. https://doi.org/10.4473/TPM28.4.5

Epifania, O. M., Anselmi, P., & Robusto, E. (2022a). Implicit social cognition through the years: The Implicit Association Test at age 21. *Psychology of Consciousness, 9*(3), 201–217. https://doi.org/10.1037/cns0000305

Epifania, O. M., Robusto, E., & Anselmi, P. (2022b). Is the performance at the Implicit Association Test sensitive to feedback presentation? A Rasch-based analysis. *Psychological Research.* Advance online publication. https://doi.org/10.1007/s00426-022-01703-w

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

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

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

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

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. *Frontiers in Psychology, 10*, Article e2483. https://doi.org/10.3389/fpsyg.2019.02483

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics, 6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*(2), 166–180. https://doi.org/10.1177/0146167204271418

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology, 44*(1), 29–45. https://doi.org/10.1348/014466604X23491

R Core Team. (2018). *R: A language and environment for statistical computing*. R Project. www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19*, 476–506.

Stefanutti, L., Robusto, E., Vianello, M., & Anselmi, P. (2013). A discrimination–association model for decomposing component processes of the Implicit Association Test. *Behavior Research Methods, 45*(2), 393–404. https://doi.org/10.3758/s13428-012-0272-3

Stefanutti, L., Robusto, E., Vianello, M., Anselmi, P., Dalla Rosa, A., & Bar-Anan, Y. (2020). Does discrimination beat association in the IAT? The discrimination-association model reconceived. *Behavior Research Methods, 52*(4), 1640–1656. https://doi.org/10.3758/s13428-019-01340-z

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204. https://doi.org/10.3102/10769986031002181

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology. General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods, 49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Appendix

Generalized Linear Model and Rasch Model

According to the Rasch model, the probability of a correct response is a function of the distance on the latent trait between respondent and stimulus characteristics:

$$P(x_{ps} = 1 \mid \theta_p, b_s) = \frac{exp(\theta_p - b_s)}{1 + exp(\theta_p - b_s)} \tag{1}$$

where $P(x_{ps} = 1)$ is the probability of respondent p to correctly respond to stimulus s, θ_p is the ability of respondent p (i.e., the amount of latent trait of respondent p) and b_s is the difficulty of

stimulus s (i.e., the amount of latent trait required by item s to obtain a correct response). The higher the value of θ_p, the higher the amount of responses correctly endorsed by respondent p. The higher the value of b_s, the lower the amount of correct responses observed on stimulus s.
In a Generalized Linear Model (GLM), the binomially distributed responses are linked to the linear combination of predictors η_{ps} by a *logit* link function. The probability of a correct response μ_{ps} given the linear combination of predictors η_{ps} is obtained as:

$$\mu_{ps} = logit^{-1}(\eta_{ps}) \frac{exp(\eta_{ps})}{1 + exp(\eta_{ps})} \tag{2}$$

Linear Model and Log-Normal Model

According to the log-normal model, the expected log-time response is a function of the distance on the latent trait between respondent and stimulus characteristics:

$$t_{ps} = \delta_s - \tau_p \tag{3}$$

where t_{ps} is the expected log-time response of respondent p to stimulus s, δ_s is the time absorbing power of stimulus s (i.e., time intensity parameter), and τ_p expresses the speed with which respondent p performs the task (i.e., speed parameter). The higher the value of δ_s, the higher the amount of time spent on stimulus s. The higher the value of τ_p, the smaller the amount of time respondent p spends on the stimuli. The expected log-time response depends on the distance between respondent and stimulus parameters.
In a Linear Model (LM), the expected log-time responses are linked to the linear combination of predictors η_{ps} by an *identity* function that follows a normal distribution:

$$t_{ps} = \beta_0 + \beta_s X_s + \beta_p X_p + \varepsilon_{ps}. \tag{4}$$

The log-normal model in Equation 3 can be equated to the LM in Equation 4, where the log-time responses are predicted by respondent and stimulus characteristics and the intercept is set at 0.

Fixed and Random Structures of the (G)LMMs

The inclusion of random effects in the linear predictors η extends (G)LMs to (Generalized) Linear Mixed-Effects Models ((G)LMM). When (G)LMMs are used to estimate the Rasch-like and log-normal parameters, the stimulus and respondent parameters are summed together (i.e., from $\theta_p - b_s$ to $\theta_p + b_s$ and from $\delta_s - \tau_p$ to $\delta_s + \tau_p$ for the Rasch and log-normal models, respectively). Consequently, the higher the value of b_s, the higher the amount of correct responses registered on stimulus s (i.e., easiness parameter), and the higher the value of τ_p, the slower respondent p is (i.e., the larger the amount of time respondent p spends on each stimulus).
Respondent and stimulus estimates of Rasch-like and log-normal models are obtained from respondent and stimulus *Best Linear Unbiased Predictors* (BLUPs, the deviation of each level of the random effects from the estimates of the fixed effects, Doran et al., 2007). Person parameters (θ_p and τ_p) derive from the random effects of the respondents, being either $\alpha_p \sim N(0, \sigma^2_{\alpha p})$ (random intercepts) or $\beta_{pc} \sim MVN(0, \Sigma_{pc})$ (random slopes in associative conditions c). Stimulus parameters

(b_s and δ_s) derive from the random effects of the stimuli, being either $\alpha_s \sim N(0, \sigma^2_{\alpha s})$ (random intercepts) or $\beta_{sc} \sim MVN(0, \Sigma_{sc})$ (random slopes in the associative conditions c). Besides the distribution of the error term (i.e., $\varepsilon \sim \text{Logistic}(0, \sigma^2)$ and $\varepsilon \sim N(0, \sigma^2)$ for the GLMMs and the LMMs, respectively), the random structures of the (G)LMMs are identical. The expected response y to each trial of the IAT i ($i \in \{1, \ldots, n\}$) of participant p ($p \in \{1, \ldots, P\}$) on stimulus s ($s \in \{1, \ldots, S\}$) in condition c ($c \in \{1, \ldots, C\}$) can be either the expected *log-odds* of the probability of a correct response (GLMMs) or the expected log-time response (LMMs). Since the fixed intercept α is set at 0 (i.e., none of the levels of the fixed slope is taken as the reference value), either the *log-odds* of a correct response for each condition (GLMMs) or the average log-time for each condition (LMMs) are estimated. The fixed structure of the models is kept constant, while the random structures vary across models.

Accuracy Models Specification

Model A1: The random intercepts of respondents and stimuli across associative conditions are specified:

$$yi = logit^{-1}\left(\alpha + \beta_c X_c + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i\right) \tag{5}$$

with $\alpha_p \sim N(0, \sigma^2_{\alpha p})$. The random structure of Model A1 provides overall respondent ability θ_p and overall stimulus easiness b_s estimates.

Model A2: The random slopes of stimuli in associative conditions and the random intercepts of respondents across associative conditions are specified:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{p[i]} + \beta_{s[i]} c_i + \varepsilon_i) \tag{6}$$

with $\beta_{sc} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{sc})$ (where $\mathbf{\Sigma}_{sc}$ is the variance-covariance matrix of the population of stimuli) and $\alpha_p \sim N(0, \sigma^2_{\alpha p})$. Model A2 provides condition–specific stimulus easiness b_{sc} and overall respondent ability θ_p estimates.

Model A3: The random slopes of respondents in associative conditions and the random intercepts of stimuli across associative conditions are specified:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{s[i]} + \beta_{p[i]} c_i + c_i) \tag{7}$$

with $\beta_{pc} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{pc})$ (where $\mathbf{\Sigma}_{pc}$ represents the variance-covariance matrix of the population of respondents) and $\alpha_s \sim N(0, \sigma^2_{\alpha s})$. The random structure of model A3 provides condition–specific respondent ability θ_{pc} and overall stimulus easiness b_s estimates.

Log-Time Models Specification

Model T1: The random intercepts of respondents and stimuli across associative conditions are specified:

$$y_i = \alpha + \beta_c X_c + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i \tag{8}$$

with $\alpha_p \sim N(0, \sigma^2_{\alpha p})$ and $\alpha_s \sim N(0, \sigma^2_{\alpha s})$. Model T1 provides overall respondent speed τ_p and overall stimulus time intensity δ_s estimates.

Model T2: The random slopes of stimuli in associative conditions and the random intercepts of respondents across associative conditions are specified:

$$y_i = \alpha + \beta_c X_c + \alpha_{p[i]} + \beta_{s[i]} c_i + \varepsilon_i \tag{9}$$

with $\beta_{sc} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{sc})$ (where $\mathbf{\Sigma}_{sc}$ is the variance-covariance matrix of the population of stimuli) and $\alpha_p \sim N(0, \sigma^2_{\alpha p})$. Model T2 provides condition–specific stimulus time intensity δ_{sc} and overall respondent speed τ_p estimates.

Model T3: The random slopes of respondents in associative conditions and the random intercepts of stimuli across associative conditions are specified:

$$y_i = \alpha + \beta_c X_c + \alpha_{s[i]} + \beta_{p[i]} c_i + \varepsilon_i \tag{10}$$

with $\beta_{pc} \sim MVN(\mathbf{0}, \mathbf{\Sigma}_{pc})$ (where $\mathbf{\Sigma}_{pc}$ represents the variance-covariance matrix of the population of respondents) and $\alpha_s \sim N(0, , \sigma^2_{\alpha s})$. This model provides condition–specific respondent speed τ_{pc} and overall stimulus time intensity δ_s estimates.