Chapter 1

Applications of (G)LMMs to IAT data

In this Chapter, two empirical applications of the modeling framework proposed in Chapter **??** are presented. In the first application, the accuracy and and log-time models for the estimation of the Rasch model and the log-normal model estimates, respectively, have been applied to an IAT for the implicit assessment of attitudes towards Black and White people (i.e., Race IAT, Section 1.1). The relationship between model estimates and the typical IAT scoring (i.e., *D* score) has been investigated as well. The second application was aimed at investigating whether the estimates obtained with accuracy and log-time models result in a better inference of the construct under investigation than that provided by the *D* score. To pursue this aim, the predictive ability of the model estimates and that of the *D* score have been compared, and an IAT for the implicit assessment of the preference for Dark and Milk chocolate was used (i.e., Chocolate IAT, Section 1.2).

The accuracy and the log-time models were fitted with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in `R` (Version 3.5.1, R Core Team, 2018). The IAT *D* scores were

computed by using the `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020).

A summary of the Rasch model estimates and the log-normal model estimates that can be obtained from the random structures of the (G)LMMs presented in Chapter **??** is reported in Table 1.1

Table 1.1: Rasch model and log-normal model estimates.

	Rasch model		Log-normal model	
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, where P, S, and C, are the number of respondents, stimuli, and conditions, respectively.

1.1 Empirical application on a Race IAT

1.1.1 Method

Participants. Sixty-five university students (F $= 49.23\%$, Age = 24.95 ± 2.09 years) voluntarily took part in the study. Participants were informed about the confidentiality of the data and asked for their consent to take part in the study. Most of them (84.62%) identified themselves as belonging to the Mediterranean ethnic group.

Materials and procedure. Participants were presented with a Race IAT. It was composed of 16 attribute stimuli, divided in 8 positive words (i.e., "love", "good", "happiness", "joy", "glory", "peace", "pleasure", "laughter") and 8 negative words (i.e., "bad", "pain", "failure", "annoying", "evil", "hate", "horrible", "terrible"), and 12 object stimuli. Object stimuli (same as in Study 2 of Nosek, Greenwald, & Banaji, 2005) were 6 Black people faces (3 male and 3 female) and 6 White People faces (3 male and 3 female). Participants were presented with 60 trials in the White-Good/Black-Bad (WGBB) condition, and 60 trials in the Black-Good/White-Bad (BGWB) one. Participants were given feedback in case of incorrect responses and were asked to correct the response to continue the experiment. They were instructed to be as accurate and fast as they could.

1.1.2 Data analysis

Data cleaning and *D* score

Exclusion criteria based on both latency and accuracy responses are applied (Greenwald, Nosek, & Banaji, 2003; Nosek, Banaji, & Greenwald, 2002). Specifically, respondents are eliminated if they show more than 25% of error responses in at least one associative condition (Nosek et al., 2002), or if they have more than 10% of the trials with a latency faster than 300ms (Greenwald et al., 2003). Trials with a latency slower than 10,000ms are eliminated as well. In applying the LMMs, the raw latencies at both correct and incorrect responses are used. The algorithm *D1* in Greenwald et al. (2003) is used for scoring the IAT. The difference is taken between the average response time in the BGWB and the WGBB condition: Positive

scores stand for a possible preference for White people over Black people.

Outfit Statistics

The fit of the data to the model is evaluated by means of Outfit statistics. Outfit statistics are a common procedure for the evaluation of the fit of each item and each respondent to the Rasch model. They are usually computed for accuracy responses, and on data where there is only one possible combination between each subject and each item. In this section, an attempt of computing Outfit statistics for the log-normal model and for the fully-crossed structure of the IAT is presented.

Rasch model. Outfit statistics on accuracy responses are computed by following a procedure close to that usually employed for their computation (e.g., Linacre, 2002). Typical Outfit computation procedures are based on the standardized residuals for only one respondent $\times$ stimulus occurrence. As already mentioned, in the IAT there are more occurrences for the combination of each respondent with each stimulus in each associative condition. Consequently, the computation is adapted to the specific data structure of the IAT.

The standardized residuals are computed as:

$$z_i = \frac{x_i - P(x_i)}{\sqrt{P(x_i = 1)P(x_i = 0)}} \tag{1.1}$$

where $P(x_i)$ is the expected probability for a correct response to each trial i of each respondent p to each stimulus s in each condition c estimated with the Rasch model, and x_i is the observed response to each trial i of each respondent p to each stimulus s in each condition c.

Normally, the Outfit statistics are computed by averaging the squared standardized residuals across respondents (stimuli Outfit) or across items (respondents Outfit), and one value for each stimulus and one for each respondent are obtained. In the IAT case, also the associative condition must be taken into account, and the number of Outfit statistics for each respondent and each stimulus depends on the random structure of the model. Specifically, if Model A2 results as the best fitting model, then condition–specific Outfit statistics u_{sc} for the stimuli are computed as:

$$u_{sc} = \frac{\sum_{p=1}^{P} z_{spc}^2}{P}, \tag{1.2}$$

where $n \in 1, \dots N$ is the number of responses obtained by the item in each associative condition. The overall Outfit statistics for the respondents u_p is obtained as:

$$u_p = \frac{\sum_{l=1}^{L} z_i^2}{L}, \tag{1.3}$$

where $l \in 1, \dots L$ is the number of trials responded by the individuals, across conditions.

Conversely, if Model A3 results as the best fitting model, then condition–specific Outfit statistics for the respondents u_{pc} are obtained as:

$$u_{pc} = \frac{\sum_{l=1}^{L} z_i^2}{L}, \tag{1.4}$$

where $l \in 1, \ldots L$ is the number of trials responded by the individuals in each condition. The overall item Outfit statistics are obtained as:

$$u_s = \frac{\sum_{n=1}^{N} z_i^2}{N}, \tag{1.5}$$

where $n \in 1, \ldots N$ is the number of responses obtained by the item across associative conditions.

Log-normal model. A similar procedure is followed for the computation of Outfit statistics on log-time responses. The difference for the computation of the residuals $z_i$is taken between the observed log-time responses to each trial t_i and the expected log-time to each trial $\bar{t}_i$ estimated by the log-normal model (Equation **??**).

If Model T2 results as the best fitting one, then condition–specific Outfit statistics u_{sc} for the stimuli can be computed following the procedure in Equation 1.2, as well as overall Outfit statistics u_p for the respondents as in Equation 1.3.

Conversely, if Model T3 results as the best fitting model, respondents' condition–specific Outfit statistics u_{pc} can be obtained as in Equation 1.4, as well as overall stimuli outfit statistics as in Equation 1.5.

For both the Outfit statistics computed on accuracy responses and log-time responses, the thresholds indicating *underfit* (i.e., the data shows a variability that the model cannot explain) or *overfit* (i.e., the data shows less variability than that expected by the model) in Linacre (2002) were used to decide on the goodness of fit of the specific respondent/stimulus

to the data. If Outfit statistics ranged between 0.50 and 2.00 (Linacre, 2002), the data were considered to have a good fit to the model. A major weight was given to respondents/stimuli showing underfit, while overfit was not considered as much problematic.

Relationship between model estimates and typical scoring

In case the best fitting model allows for the multidimensionality of the associative conditions on the respondents, differential measures (either *ability-differential* or *speed-differential*) are computed. The relationship between the estimates of the Rasch model, those of the log-normal model, their potential differential measures, and the typical IAT *D* score are investigated with Person's correlation and by regressing the linear combination of the respondents' estimates on the *D* score. In case of differential measures, both differential measures and their linear combination are regressed on the *D* score in separate models. This is done to determine the actual weight of each condition–specific estimate on the final *D* score. Backward deletion is used for investigating the predictor(s) that explains the higher amount of variance of the *D* score.

1.1.3 Results

No participants or trials were eliminated grounding on the response time exclusion criteria 1.1.2. Three participants were excluded because of the accuracy deletion criterion (Nosek et al., 2002). The sample was finally composed by 62 participants (F $= 48.39\%$, Age = 24.92 ± 2.11 years).

The overall average response time was 815.06 ms (*sd* $= 423.20$, *skewness* $= 3.82$, *kurtosis*

$= 33.87$), while the average response time was 667.11 ms in the WGBB condition (*sd* $=$ 294.06, *skewness* $= 4.64$, *kurtosis* $= 44.60$) and 943.01 ms (*sd* = 488.89, *skewness* $= 3.45$, *kurtosis* $= 29.05$) in the BGWB one. After the log-transformation of the response latencies (expressed in second), the overall average response time was -0.29 log-seconds (*sd* $= 0.40$, *skewness* $= 0.72$, *kurtosis* $= 3.88$), the average response time was -0.43 log-seconds in the WGBB condition (*sd* $= 0.31$, *skewness* $= 1.26$, *kurtosis* $= 3.73$), and the average response time was -0.15 log-seconds in the BGWB condition (*sd* $= 0.42$, *skewness* $= 0.24$, *kurtosis* $= 5.09$).

Rasch models

The accuracy models in Table 1.1 were applied to the Race IAT. Concerning AIC, Log-Likelihood, and Deviance, Model A2 (AIC = 3784.43, Log-Likelihood = -1886.21, Deviance = 3722.43) performed better than Model A3 (AIC = 3786.51, Log-Likelihood = -1887.26, Deviance = 3774.51) and Model A1 (AIC = 3785.87, Log-Likelihood = -1888.93, Deviance = 3777.87). However, the latter one showed the lowest BIC value (3813.53, 3825.91, 3828.00, BIC values for Model A1, A2, and A3, respectively). Model A2 was chosen. This model provided overall participants ability parameters θ_p and condition–specific stimuli easiness parameters (b_{WGBB} and b_{BGWB}). The estimates of the fixed effects of Model A2 indicated a higher probability of correct response in the WGBB condition (*log-odds* $= 3.45$, *SE* $= 0.12$) than in the BGWB condition (*log-odds* $= 2.07$, *SE* $= 0.11$). Between–participants variability was 0.17. Between–stimuli variability in the WGBB condition ($\sigma^2 = 0.08$) was lower than the between–stimuli variability in the BGWB condition ($\sigma^2 = 0.15$). The correlation

between stimuli variability in the two conditions was moderate ($r = .34$).

Outfit statistics of the respondents ranged between 0.04 and 1.85 ($M = 0.92 \pm 0.33$). Seven respondents showed Outfit statistics below 0.50, and they were retained in the analysis.

All stimuli showed appropriate Outfit values in condition BGWB ($M = 0.92 \pm 0.12$, $Min = 0.69$, $Max = 1.08$). Outfit statistics in condition WGBB ($M = 0.94 \pm 0.40$, $Min = 0.25$, $Max = 1.71$) highlighted four stimuli with Outfit values below 0.50, but they were retained in the analysis. Stimuli easiness parameters for each condition resulting from Model A2 are reported in Table 1.2.

Table 1.2: Stimuli condition–specific easiness parameters (b_{sc}) and overall time intensity parameters (δ_s) - Race IAT

	b_{WGBB}	b_{BGWB}	$b_{\text{WGBB}} - b_{\text{WGBB}}$	δ_s		b_{WGBB}	b_{BGWB}	$b_{\text{WGBB}} - b_{\text{WGBB}}$	δ_s
Good attributes					*Bad* attributes				
joy	3.53	1.69	1.85	0.02	evil	3.19	1.37	1.82	−0.01
happiness	3.48	1.67	1.81	0.01	horrible	3.56	1.77	1.79	0.05
pleasure	3.29	1.60	1.69	0.05	bad	3.11	1.58	1.53	0.03
peace	3.32	1.73	1.59	0.01	terrible	3.34	1.81	1.52	0.01
good	3.54	1.95	1.59	0.01	hate	3.34	1.85	1.50	0.01
laughter	3.54	2.03	1.52	0.09	failure	3.43	2.06	1.38	0.05
love	3.48	1.99	1.49	0.01	annoying	3.07	1.87	1.20	0.09
glory	3.42	1.99	1.43	0.08	pain	3.21	2.02	1.19	0.10
M (*SD*)	3.45 (0.09)	1.83 (0.16)	1.62 (0.15)	0.03 (0.04)		3.28 (0.15)	1.79 (0.21)	1.49 (0.22)	0.04 (0.04)
White people faces					*Black* people faces				
wm3	3.61	2.04	1.57	−0.05	bm2	3.61	2.32	1.30	−0.08
wf3	3.66	2.29	1.36	−0.05	bf2	3.56	2.33	1.23	−0.06
wf2	3.59	2.46	1.12	−0.03	bf1	3.56	2.36	1.20	−0.04
wm2	3.48	2.44	1.04	0.03	bm1	3.52	2.42	1.10	−0.10
wf1	3.59	2.57	1.02	−0.05	bm3	3.58	2.51	1.07	−0.09
wm1	3.28	2.28	1.01	−0.02	bf3	3.36	2.47	0.89	−0.05
M (*SD*)	3.54 (0.14)	2.35 (0.17)	1.19 (0.21)	−0.03 (0.03)		3.53 (0.09)	2.40 (0.07)	1.13 (0.13)	−0.07 (0.02)

Note: "wf": White person female face; "wm": White person male face; "bf": Black person female face; "bm": Black person male face; WGBB: White-Good/Black-Bad condition; BGWB: Black-Good/White-Bad condition. Rows are ordered by decreasing values of $b_{\text{WGBB}} - b_{\text{WGBB}}$

Overall, the IAT stimuli tended to be easy stimuli. Stimuli tended to be easier in the WGBB condition ($M = 3.44 \pm 0.16$) than in the BGWB condition ($M = 2.05 \pm 0.33$, $t(39) = 19.89$, $p < .001$). The category of the stimuli was used to predict the difference between the condition–specific easiness estimates, and the intercept was removed so that none of the categories was taken as a reference for the others. A significant effect of the catgeory of stimuli was found ($F(4, 24) = 359.87$. $p < .001$). The categories of stimuli that gave the highest contribution to the IAT effect were the evaluative dimensions ($B_{\text{Bad}} = 1.49$, *se* $= 0.07$, $p < .001$, and $B_{\text{Good}} = 1.62$, *se* $= 0.07$, $p < .001$). The target objects categories gave a lower contribution to the IAT effect ($B_{\text{Black}} = 1.13$, *se* $= 0.08$, $p < .001$, $B_{\text{White}} = 1.18$, *se* $= 0.08$, $p < .001$). The stimuli that gave the highest contribution to the IAT effect were *joy* and *happiness* (category *Good*), *evil* and *horrible* (category *Bad*), *wm3* and *wf3* (category *White*), and *bm2* and *bf2* (category *Black*). The stimuli that gave the lowest contribution to the IAT effect were *love* and *glory* (category *Good*), *annoying* and *pain* (category *Bad*), *wf1* and *wm1* (category *White*) and *bm3* and *bf3* (category *Black*).

Log-normal models

The log-time models in Table 1.1 were applied to the Race IAT. Model T2 produced aberrant estimates (i.e., correlation between the stimuli random slopes equal to 1). Model T3 (AIC = 4399.66, BIC = 4448.06, Log-Likelihood = -2192.83, Deviance = 4385.66) performed better than Model T1 (AIC = 4762.63, BIC = 4797.20, Log-Likelihood = -2376.32, Deviance = 4752.63). Model T3 was chosen. This model provided condition–specific participants' speed parameters (τ_{WGBB} and τ_{BGWB}) and overall stimuli time intensity parameters δ_j. Respondents'

Outfit statistics showed a good fit for all respondents in both associative conditions ($M = 0.98 \pm 0.01$, $Min = 0.98$, $Max = 0.99$ for the BGWB condition, and $M = 0.99 \pm 0.01$, $Min = 0.98$, $Max = 1.03$ for the WGBB condition). Overall Outfit statistics indicated a good fit for all stimuli ($M = 1.00 \pm 0.16$, $Min = 0.77$, $Max = 1.33$).

Responses in the WGBB condition were faster ($B = -0.43$, $SE = 0.02$) than responses in the BGWB condition ($B = -0.15$, $SE = 0.03$). The between-stimuli variability was particularly low ($\sigma^2 = 0.003$), while the between–participants variability was slightly higher in the BGWB condition ($\sigma^2 = 0.05$) than in the WGBB one ($\sigma^2 = 0.02$). The correlation between respondents' variability in the two conditions was strong ($r = .63$).

Stimuli time intensity parameters δ_s obtained from Model T3 are reported in Table 1.2. A significant effect of the categories of stimuli was found on the time intensity estimates ($F(4, 24) = 11.77$, $p < .001$). The exemplars of both evaluative dimensions tended to require a high amount of time for getting a response ($B_{\text{Bad}} = 0.04$, $se = 0.01$, $p < .001$, and $B_{\text{Good}} = 0.03$, $se = 0.01$, $p = .01$). The exemplars of the category *Black* were the stimuli requiring the least time for getting a response ($B = -0.07$, $se = 0.01$, $p = 0.01$), immediately followed by the exemplars of the category *White* ($B = -0.03$, $se = 0.01$, $p = 0.04$).

Three of the positive attribute stimuli (*pleasure*, *glory*, *laughter*) showed time intensity estimates higher than the estimates of the stimuli belonging to the same category. Also three negative attributes (*failure*, *annoying*, *pain*) showed a higher time intensity estimates than the other negative attributes. Object stimuli tended to have similar time intensity estimates.

Relationship between model estimates and typical scoring

A *speed-differential* measure was computed as the difference between respondents' speed estimates in the BGWB condition and those in the WGBB condition. Negative values indicated a respondent with a higher speed in the BGWB condition than in the WGBB condition. Pearson's correlations were computed between participants' ability, condition–specific speed parameters, and *speed-differential*. Participants' ability poorly and positively correlated with speed in the BGWB condition ($r = .13$, $p = .32$), and it poorly and negatively correlated with the *speed-differential* ($r = -.14$, $p = .28$), although these correlations were not significant. Ability moderately correlated with speed in the WGBB condition ($r = .32$, $p = .01$).

Respondents' ability and *speed-differential* were regressed on the *D* score. Backward deletion kept both the predictors in the model, which accounted for about 70% of the total variance (*Adjusted* $R^2 = .78$, $F(2, 59) = 106.3$, $p < .001$). *Speed-differential* strongly and positively predicted the *D* score ($B = 1.93$, $t(59) = 13.88$, $p < .001$), whereas ability negatively predicted the *D* score ($B = -0.18$, $t(59) = -2.48$, $p = .016$).

To better understand the specific contribution of the speed of each associative condition, a model including the linear combination of the ability estimate, the speed estimate in the WGBB condition, and the speed estimate in the BGWB condition was specified as well. Backward deletion kept all predictors in the model, which accounted for almost the 80% of the total variance (*Adjusted* $R^2 = .79$, $F(3, 58) = 76.46$, $p < .001$). The speed estimate in the WGBB condition negatively predicted the *D* score($B = –2.22$, $t(58) = –11.43$, $p < .001$), while the speed in the BGWB condition positively predicted it ($B = 1.92$, $t(58) = 14.16$,

$p < .001$). Despite the ability estimate remained in the model, its contribution was no longer significant ($B = –0.13$, $t(58) = –1.76$, $p = .08$).

1.1.4 Final remarks

The fine-grained analysis at the level of the stimuli allowed for the investigation of the stimuli representativeness of the category to which they belong, as well as of their contribution to the IAT effect. Besides leading to a deeper understanding of the IAT effect and hence of the measure itself, this information can also be exploited for the design of brief, but still highly informative IATs by selecting the most informative and prototypical stimuli.

The selection of a smaller but highly informative pool of stimuli is also expected to lower the across-trial variability. Consequently, also the *D* score should result in a more reliable estimate of the implicit construct under investigation. The information at the stimuli level can inform about the implicit evaluative associations driving the performance. In this instance, the evaluative dimensions *Good* and *Bad* were the stimuli categories showing the highest difference between the associative conditions. Both stimuli categories were easier in the WGBB condition than in the BGWB condition. This implies that *Good* exemplars were more easily sorted when their category shared the response key with category *White* than when it shared the response key with category *Black*. Similarly, *Bad* attributes were more easily sorted when their category shared the response key with category *Black* than when it shared the response key with category *White*. This result is in line with the positive primacy effect in Anselmi, Vianello, and Robusto (2011).

The overall ability estimates indicate a low within–respondents between–conditions vari-

ability in the accuracy performance of the respondents. This implies that their accuracy performance did not change according to the associative conditions. Conversely, the condition–specific speed estimates indicate that respondent' speed performance did vary between conditions. Taken together, these results show that the respondents tend to slow down in the condition against their own automatically activated association to keep their accuracy performance unaltered. Evidence for this effect has already been found in the literature (speed-accuracy trade-off, Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007), and it is indeed a common (and expected) phenomenon in speeded computerized tasks (van der Linden, 2006, 2009).

Moreover, the respondents' speed and ability estimates allowed for a deeper understanding of the IAT effect as it is expressed by the D score. Ability was poorly related with the D score, while condition–specific speed estimates pinpointed the higher contribution of the speed in the WGBB condition than of that of the speed in the BGWB. This result is consistent with what already highlighted by the results on the contribution of each stimulus to the IAT effect.

In this application, the relationship between model estimates and external criteria, such as the explicit assessment of the same construct or behavioral outcomes was not investigated. Therefore, conclusions on the validity of the model estimates should be interpreted with caution and further evidence is needed.

1.2 Empirical application on a Chocolate IAT

Since the estimates obtained from the modeling framework that has been proposed should not be influenced by unwanted error variance due to the non-independence of the observations, they are supposed to be more reliable than the *D* score. As such, it can be speculated that these estimates provide a better inference of the construct under investigation. The better inference provided by this modeling approach is expected to lead to a more accurate prediction of behavioral outcome, as well as to show a stronger relationship with explicit measures of the same construct. However, this hypothesis was not tested in the previous application.

This study was aimed at testing this hypothesis by comparing the predictive ability of the Rasch and log-normal model estimates and that of the *D* score in respect to a dichotomous behavioral choice. Moreover, previous studies already highlighted the importance of stimuli representativeness for a correct functioning of the IAT. Specifically, it has been suggested that it is better to use a smaller but highly representative set of stimuli than a larger one including also poorly representative stimuli (Nosek et al., 2005). Clearly, a large pool of poorly representative stimuli results in both high variability at the stimuli level and scarce information provided by the stimuli. A high across-trial variability due to stimuli heterogeneity deeply affects the *D* score computation (Wolsiefer, Westfall, & Judd, 2017). Conversely, by using a smaller set of highly prototypical and representative stimuli, the across-trial variability should be reduced. As such, the *D* score computation is less affected by error variance components due to the across-trial variability, resulting in a more accurate inference of the construct under investigation. Following this line of reasoning, the *D* score should result in a better predictive

ability of behavioral outcomes when it is computed on a smaller data set composed of highly representative stimuli than when it is computed on either the entire set of stimuli or a smaller data set composed of poorly representative stimuli.

To test this hypothesis, the information provided by the Rasch model and the log-normal model at the stimuli level was exploited to select the most informative and the least informative stimuli for each category. Smaller data set were obtained by exploiting this information.

1.2.1 Method

Participants. Seventy-six university students (F $= 71.05\%$, Age = 24.02 ± 2.88 years) volunteered to take part in the study. They were informed about the confidentiality of the data and they were asked for their consent to take part in the study.

Materials and procedure. The Chocolate IAT used the same stimuli described in Epifania, Robusto, and Anselmi (2020). Specifically, twenty-six attribute stimuli (13 *Good* exemplars and 13 *Bad* exemplars) and fourteen chocolate images (7 *Dark* chocolate and 7 Milk chocolate) were used. Images were obtained from the same starting images, which were appropriately modified to represent either *Dark* or *Milk* chocolate.

Respondents were presented with 60 trials in the Dark-Good/Milk-Bad (DGMB) condition, and 60 trials in the Milk-Good/Dark-Bad (MGDB) condition. No feedback was given in case of incorrect responses. Respondents were asked to be as fast and as accurate as they could.

Respondents' explicit chocolate preferences were investigated with two items (i.e., *"How*

much do you like Milk chocolate?" and *"How much do you like Dark chocolate?"*) evaluated on a 6 points Likert-type scale ("0 - Not at all", "5 - Very much"). They were also asked about their food habits and behaviors through a 6-item scale (Cronbach's $\alpha = 0.80$, example item *"I am usually on a diet"*), rated on a 4-point agreement Likert-type scale ("1 - Strongly Disagree", "4 - Strongly agree"). Higher scores indicated higher care about food habits. At the end of the experiment, participants were invited to choose between a free dark or milk chocolate bar as a reward for their participation. The experimenter registered their choice after they left the laboratory. Participants performed the experiment individually in a laboratory setting.

1.2.2 Data analysis

Data cleaning and *D* score computation

Exclusion criteria based on both accuracy (Nosek et al., 2002) and time responses (Greenwald et al., 2003) are applied (see Section 1.1.2). The *D4* algorithm in Greenwald et al. (2003) is used to score the IAT. The difference is taken between the average response time in the MGDB condition and that in the DGMB condition. Therefore, positive scores stand for a possible preference for Dark chocolate over Milk chocolate. LMMs are applied to the raw log-time responses of both correct and incorrect responses, without any penalties.

Oufit statistics

The same procedure for computing Outfit statistics and the thresholds for interpreting them, are as those used in the previous study.

Relationship between model estimates, typical scoring, and explicit measures

The relationships between the Rasch model respondents' estimates, the log-normal model respondents' estimates, the D score, and explicit chocolate evaluations are investigated by computing Pearson's correlations. If the best fitting models allow for the multidimensionality at the respondents level, so that condition–specific estimates are obtained, differential measures are computed, and their relationship with the above mentioned variables is investigated as well.

Predictive ability of a behavioral outcome

To investigate the predictive ability of the Rasch model estimates, that of the log-normal estimates, and of the D score, separate logistic regressions are specified. Dark chocolate choice (DCC) is labeled as 0, and Milk chocolate choice (MCC) is labeled as 1.

If the best fitting model for the Rasch model or log-normal model allow for the multidimensionality at the respondents' level, hence condition–specific respondents' estimates are obtained, then differential measures, *ability-differential* or *speed-differential* are computed, and used for the prediction. In such cases, both the predictive ability of the differential measures and that of the linear combination of their single components are investigated. The predictive ability of the single components of the D score is investigated as well. The single

components of the *D* score are the average response times (computed on the already corrected response times) in each associative condition.

Predictive ability of the reduced data sets. The information at the stimuli level can be used to select the most and least informative stimuli for each stimuli categories. The two most informative stimuli for each category, and the least informative stimuli for each category, are selected to create smaller data sets, an highly informative one ("Best") and a lowly informative one ("Worst"). In both cases, the stimuli pool is composed of 8 stimuli. A *D* score is computed on each of the newly obtained data set, and it is used for predicting the choice. Their performance is compared with that of the *D* score computed on the entire data set. All starting models include food habits, and relevant predictors are selected with backward deletion. Model general accuracy (i.e., percentage of choices correctly identified by the model), model DCC accuracy (i.e., percentage of DCCs correctly identified by the model), and model MCC accuracy (i.e., percentage of MCCs correctly identified by the model) are used as criteria to establish the predictors best accounting for the actual choice. Nagelkerke's R^2 (Nagelkerke, 1991) is used as Pseudo R^2.

1.2.3 Results

One trial was eliminated because of a latency higher than 10,000 ms. Two participants were eliminated grounding on the accuracy elimination criterion (Nosek et al., 2002). The final sample was composed of 74 participants (F $= 71.62\%$, Age = 24.08 ± 2.88 years). Milk chocolate was chosen by 41.90% of the participants.

The overall average response time was 858.99 ms (*sd* $= 503.08$, *skewness* $= 3.85$, *kurtosis* $= 29.34$). The average response time was 973.80 ms in the DGMB condition (*sd* $= 557.08$, *skewness* $= 3.07$, *kurtosis* $= 16.90$) and 744.20 ms (*sd* = 411.75, *skewness* $= 5.75$, *kurtosis* $= 71.07$) in the MGDB one. After the log-transformation of the response latencies (expressed in second), the overall average response time was -0.26 log-seconds (*sd* $= 0.43$, *skewness* $= 1.00$, *kurtosis* $= 1.48$), the average response time was -0.14 log-seconds in the DGMB condition (*sd* $= 0.45$, *skewness* $= 0.72$, *kurtosis* $= 0.93$), and the average response time was -0.38 log-second in the MGDB condition (*sd* $= 0.37$, *skewness* $= 1.38$, *kurtosis* $= 3.17$).

Rasch models

The accuracy models presented Table 1.1 were applied to the Chocolate IAT. Model A3 failed to converge, while Model A2 (AIC = 3625.58, Log-Likelihood = -1806.79, Deviance = 3613.58) performed better than Model A1 (AIC = 3627.71, Log-Likelihood = -1809.85, Deviance = 3619.71). Model A1 showed a lower value of BIC than Model A2 (3656.07, 3668.13 for Model A1 and Model A2, respectively). Model A2 was chosen. The model resulted in the estimation of overall respondents' ability θ_p and condition–specific easiness parameters (b_{MGDB} and b_{DGMB}).

A higher probability of a correct response was found in the MGDB condition (*log-odds* $= 3.67$, *SE* $= 0.14$) than in the DGMB one (*log-odds* $= 2.61$, *SE* $= 0.10$). Between–respondents variability was high ($\sigma^2 = 0.33$). Between-stimuli variability was higher in the MGDB condition ($\sigma^2 = 0.21$) than in the DGMB condition ($\sigma^2 = 0.01$). The variability of the stimuli in the two conditions were weakly correlated ($r = .20$).

Respondents' Outfit statistics ranged between 0.02 and 1.53 ($M = 0.87 \pm 0.31$). Five respondents showed Outfit values below 0.50, but they were retained in the analysis.

Four stimuli in the DGMB condition showed Outfit statistics below 0.50 ($M = 0.89 \pm 0.30$, $Min = 0.31$, $Max = 1.45$) and ten stimuli in the MGDB condition showed Outfit statistics below 0.50 ($M = 0.85 \pm 0.44$, $Min = 0.02$, $Max = 1.87$). All stimuli were retained in the analysis.

Stimuli easiness parameters are reported in Table 1.3.

Table 1.3: Stimuli condition–specific easiness parameters (b_{sc}) and overall time intensity parameters (δ_s) - Chocolate IAT

	b_{DGMB}	b_{MGDB}	$b_{\text{DGMB}} - b_{\text{DGMB}}$	δ_s		b_{DGMB}	b_{MGDB}	$b_{\text{DGMB}} - b_{\text{DGMB}}$	δ_s
Good attributes					*Bad* attributes				
joy	2.62	4.02	−1.40	0.01	hate	2.59	3.85	−1.26	0.01
happiness	2.64	4.03	−1.39	0.02	failure	2.68	3.93	−1.25	0.07
pleasure	2.56	3.70	−1.15	0.01	terrible	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	2.64	3.59	−0.95	0.01	pain	2.58	3.23	−0.65	0.05
beauty	2.61	3.46	−0.85	0.02	annoying	2.58	3.05	−0.47	0.08
wonderful	2.62	3.45	−0.83	0.09	agony	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)		2.61 (0.03)	3.53 (0.41)	−0.92 (0.40)	0.05 (0.03)
Dark Chocolate					*Milk* Chocolate				
Dark5	2.56	3.94	−1.38	−0.12	Milk3	2.60	3.95	−1.35	−0.04
Dark2	2.60	3.82	−1.23	−0.11	Milk6	2.66	3.99	−1.33	−0.04
Dark6	2.55	3.72	−1.16	−0.10	Milk4	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	2.58	3.53	−0.95	−0.08	Milk5	2.62	3.64	−1.02	−0.05
Dark7	2.58	3.41	−0.83	−0.07	Milk1	2.62	3.62	−1.01	−0.03
Dark1	2.49	3.27	−0.78	−0.11	Milk7	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)		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; Difference: Difference between DGMB and MGDB condition. Rows are ordered by absolute decreasing values of $b_{\text{DGMB}} - b_{\text{DGMB}}$. The units of the easiness estimates are the *log-odds*, the units of the time intensity estimates are the log-seconds.

Irrespective of the category to which they belong, stimuli tended to be easier in the MGDB condition ($M = 3.63 \pm 0.29$) than in the DGMB one ($M = 2.60 \pm 0.04$, $t(40) = -21.97$. $p < .001$). A significant effect of the categories of the stimuli was found on the difference in the easiness estimates between the associative conditions. The exemplars of the category *Milk* ($B = -1.13$, *se* $= 0.11$, $p < .001$) and those of the category *Good* ($B = -1.09$, *se* $= 0.08$, $p < .001$) were the stimuli that gave the highest contribution to the IAT effect. The category of stimuli that gave the least contribution to the IAT effect was the category *Bad* ($B = -0.92$, *se* $= 0.07$, $p < .001$), followed by the contribution given by the category *Dark* ($B = -1.05$, *se* $= 0.11$, $p < .001$).

The difference between condition–specific easiness estimates is used for pinpointing the stimuli giving the highest contribution to the IAT effect as well as the stimuli giving the least contribution. According to the condition–specific easiness difference, the stimuli giving the highest contribution to the IAT effect were *joy*, *happiness*, and *pleasure* (category *Good*), *hate*, *failure*, and *terrible* (category *Bad*), *Dark5*, *Dark2*, and *Dark6* (category *Dark*)), and *Milk6*, *Milk3*, and *Milk4* (category *Milk*). The three stimuli that gave the least contribution to the IAT effect were *beauty*, *wonderful*, and *excellent* (category *Good*), *annoying*, *agony*, and *pain* (category *Bad*), *Dark 1*, *Dark 7*, and *Dark 3* (category *Dark*), and *Milk 1*, *Milk 7*, and *Milk 5* (category *Milk*).

Log-normal models

The log-time models presented Table 1.1 were applied to the Chocolate IAT. Model T2 produced aberrant estimates. Model T3 (AIC = 7159.23, BIC = 7208.87, Log-Likelihood =

-3572.62, Deviance = 7145.23) performed better than model T1 (AIC = 7856.45, BIC = 7891.91, Log-Likelihood = -3923.23, Deviance = 7846.45). Thus, model T3 was chosen. The model resulted in overall stimuli time intensity parameters δ_k and respondents' condition–specific speed parameters (τ_{MGDB} and τ_{MGDB}). Responses tended to be faster in the MGDB condition ($B = -0.36$, *SE* $= 0.02$) than in the DGMB condition ($B = -0.12$, *SE* $= 0.03$). Between-stimuli variability was extremely low ($\sigma^2 = 0.004$). Between-participants variability was higher in the DGMB condition ($\sigma^2 = 0.05$) than in the MGDB one ($\sigma^2 = 0.03$). The correlation between participants' slopes in the two conditions was moderate ($r = .40$).

Respondents' Outfit statistics showed a good fit for all respondents in both associative conditions (*M* $= 0.98 \pm 0.01$, *Min* $= 0.97$, *Max* $= 1.00$ for the DGMB condition, and *M* $= 0.99 \pm 0.01$, *Min* $= 0.98$, *Max* $= 0.99$ for the MGDB condition). Outfit statistics indicated a good fit for all stimuli (*M* $= 0.99 \pm 0.12$, *Min* $= 0.73$, *Max* $= 1.28$).

Stimuli time intensity parameters δ_k are reported in Table 1.3. A significant effect of the categories of the stimuli was found on the stimuli time intensity estimate ($F(4, 36) = 37.41$, $p < .001$). The exemplars of both target objects categories required the least amount of time for getting a response ($B_{\text{Dark}} = -0.09$, *se* $= 0.01$, $p < .001$, and $B_{\text{Milk}} = -0.04$, *se* $= 0.01$, $p < .001$). The exemplars of the category *Bad* were the stimuli that required the highest amount of time for getting a response ($B = 0.05$, *se* $= 0.01$, $p < .001$), followed by those belonging to the category *Good* ($B = 0.03$, *se* $= 0.01$, $p < .001$). It was possible to identify stimuli with time intensity estimates far away from the time intensity estimates of the stimuli belonging to the same category. For instance, stimulus *heaven* was the stimulus requiring

more time within the *Good* category.

Relationship between model estimates, typical scoring, and explicit measures

A *speed-differential* was computed as the difference between respondents' speed estimates in the MGDB condition and those in the DGMB condition. Positive values indicated a higher speed in the DGMB condition than in the opposite one. Pearson's correlation was computed between participants' ability, condition–specific speed parameters, and *speed-differential*.

Results of Pearson's correlations computed between respondents' explicit preference for Milk and Dark chocolate, *D* scores, ability estimates, condition–specific speed estimates, and *speed-differential* are reported in Table 1.4.

Table 1.4: Correlation between model estimates, explicit measures, and *D* scores.

	1	2	3	4	5	6	7
1 - Explicit Milk							
2 - Explicit Dark	-0.51^{***}						
3 - *D*	-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: *** $p < .001$, ** $p < .01$, *D*: IAT *D* score, τ: speed estimate, θ: Ability estimate, DGMB: Dark-Good/Milk-Bad condition, MGDB: Milk-Good/Dark-Bad condition, *Speed-differential*: $\tau_{\text{MGDB}} - \tau_{\text{DGMB}}$.

Explicit chocolate evaluations strongly and negatively correlated between each other. Each explicit chocolate evaluation strongly correlated with the *D* score, consistently with the direction with which it was computed. The more Milk (Dark) chocolate was positively

evaluated on the explicit measure, the higher the speed in the condition where Milk (Dark) was associated with *Good* attributes, as it is pointed out by the negative correlations between explicit evaluations and condition–specific speed estimates.

The *D* score strongly and negatively correlated with the speed estimates in the DGMB condition. The *D* score showed a positive, although slightly weaker, correlation with speed estimates in MGDB condition.

The *speed-differential* negatively correlated with Milk chocolate explicit evaluation and positively correlated with Dark chocolate explicit evaluation. The *D* score and speed-differential strongly and positively correlated between each other. The *speed-differential* moderately and negatively correlated with the speed estimate of the DGMB and it strongly and positively correlated with speed estimate in the MGDB condition.

Taken together, these results suggest that the IAT effect is mostly driven by the speed in the DGMB condition.

Predictive ability of a behavioral outcome

The information provided by the difference between the condition–specific easiness estimates (see comment to Table 1.3 in Section 1.2.3) was used to create two smaller data sets. The starting data set was composed of 8,879 observations. In a first data set ("Best"), only the responses to the stimuli giving the highest contribution to the IAT effect were selected. This data set resulted in a total of 2,941 observations, ranging from a minimum of 38 observations to a maximum of 41 observations per participant. In a second data set ("Worst"), only the responses to the stimuli giving the lowest contribution to the IAT effect were selected. This

data set resulted in a total number of 2,587 observations, with a minimum of 38 observations and a maximum of 42 observations per participant. The *D4* algorithm was computed for both the Best data set and the Worst data set. In both cases, the data set was reduced to about $1/3$ of the total number of observations.

Both the predictive ability of the differential measures (i.e., *D* score and *speed-differential*) and that of their single components (i.e., M_{MGDB} and M_{DGMB} for *D* score, τ_{MGDB} and τ_{DGMB} for *speed-differential*) were investigated. Eight logistic regression models were specified, including one of the relevant predictors (or their linear combination) at the time.

Results of backward deletion are reported in Table 1.5.

The *Speed-differential* showed a slightly better general accuracy, due to a small gain in DCC accuracy, than the *D* score. Interestingly, the *D* scores computed on both Best and Worst data sets showed a better general accuracy than both the *D* score computed on the entire data set and *speed-differential*. The *D* score computed on the Best data set showed the highest general accuracy, resulting from a gain on both DCC accuracy and MCC accuracy. It also explained the highest proportion of variance. Conversely, *D* computed on the Worst data set explained the lowest proportion of variance.

All single components of the *D* score showed *log-odds* for the choice prediction near zero, regardless of the data set on which they were computed. Therefore, they did not add anything to the prediction provided by the intercept (i.e., the expected *log-odds* of the probability of choosing Milk chocolate). The single components computed on the entire data set and the Best data set showed the same general, DCC, and MCC accuracy. The single components computed on the Worst data set showed a slightly lower general accuracy, due to a loss in

Table 1.5: Choice prediction results for the differential measures and their Single components.

Predictors	*log-odds*	*se*	*Nagelkerke* R^2	*Gen*	*DCC*	*MCC*
			Differential measures			
Intercept	-1.65^{**}	0.51	0.26	0.66	0.70	0.61
D score	-2.03^{***}	0.60				
Intercept	-1.65^{***}	0.48	0.26	0.68	0.72	0.61
Speed-differential	-5.02^{***}	1.43				
Intercept	-1.76^{***}	0.52	0.30	0.70	0.74	0.65
D score (Best)	-2.07^{***}	0.58				
Intercept	-1.23^{***}	0.42	0.18	0.69	0.72	0.65
D score (Worst)	-1.40^{***}	0.47				
			Single components			
Intercept	-0.23	1.36	0.27	0.65	0.74	0.52
M_{DGMB}	0.00^{**}	0.01				
M_{MGDB}	-0.01^{**}	0.01				
Intercept	-2.05^{*}	0.74	0.27	0.72	0.74	0.68
τ_{DGMB}	4.73^{***}	1.48				
τ_{MGDB}	-5.99^{***}	1.98				
Intercept	-0.17	1.61	0.30	0.65	0.74	0.52
M_{DGMB} (Best)	0.00^{***}	0.01				
M_{MGDB} (Best)	-0.01^{*}	0.01				
Intercept	0.61	1.23	0.16	0.64	0.77	0.45
M_{DGMB} (Worst)	0.00^{*}	0.01				
M_{MGDB} (Worst)	0.00^{*}	0.01				

Note: $^{***}\, p < .001$, $^{**}\, p < .01$, $^{*}\, p < .05$, *log-odds*: Log-odds of the probability of choosing Milk chocolate, Best: Highly contributing stimuli data set, Worst: Lowly contributing stimuli data set, τ: Speed estimate, *speed-differential*: differential measure computed as $\tau_{\text{MGDB}} - \tau_{\text{DGMB}}$, DGMB: Dark-Good/Milk-Bad associative condition, MGDB: Milk-Good/Dark-Bad condition.

MCC accuracy, although it was counterbalanced by a gain in DCC accuracy.

Condition–specific speed estimates showed the highest general accuracy, due to a gain in MCC accuracy.

1.2.4 Final remarks

Results of the study reported in this section corroborate the higher reliability of the estimates obtained with statistical models able to account for IAT error variance and its random structure than that of the typical scoring method of the IAT. The information at the stimuli level obtained can be used to reduce the across-trial variability, hence resulting in a better IAT measure as expressed by the *D* score.

Results on respondents' speed (affected by the associative conditions) and accuracy (not affected by the associative condition) were in line with both the results in the previous section and the speed-accuracy trade-off in Klauer et al. (2007).

The condition–specific estimates highlighted that the IAT effect was mostly driven by *Good* attributes and *Milk* chocolate exemplars. As such, it can be speculated that it is more the liking for milk chocolate than the dislike for dark chocolate that drives the IAT effect. Consistently with this result, the magnitude of the correlations between condition–specific speed estimates and both the *D* score and the *speed-differential* suggest that speed in the MGDB condition has a major influence on the final score.

Also the overall time intensity estimates provided useful information on the stimuli functioning. Time intensity estimates highlighted different processing times both between stimuli categories (i.e., images require less time for getting a response than attributes) and within the

same stimuli category (i.e., the stimuli showing a time intensity estimate far away from the estimates of the stimuli belonging to the same category).

Finally, by selecting the stimuli that gave the highest contribution to the IAT effect, a *D* score resulting in a better prediction of the choice can be obtained. These models allow for highlighting the most representative and prototypical exemplars of each category. As such, it is possible to select the two best working stimuli to design valid and highly informative IATs, in line with what suggested by Nosek et al. (2005). Given that a lower number of stimuli is presented to the respondents, the number of trials can be reduced without losing information. The administration time of the IAT can hence be reduced. The reduction of the administration time might be useful both in an experimental laboratory setting and in online experiment. In a laboratory setting, the experimenter can control potential artifacts disrupting the administration and hence the performance of the respondent (e.g., someone enters the room, the respondent gets distracted). Nonetheless, having an IAT that requires less time gives the possibility of administering multiple measures and, most importantly, of not tiring out the respondent. In an online setting, a shorter administration time is definitely a good incentive for the participation, and might prevent respondents' withdrawal out of boredom and/or tiredness.

References

Anselmi, P., Vianello, M., & Robusto, E. (2011). Positive associations primacy in the IAT : A Many-Facet Rasch Measurement analysis. *Experimental Psychology*, *58*(5), 376–384. doi: 10.1027/1618-3169/a000106

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

Epifania, O. M., Anselmi, P., & Robusto, E. (2020). Implicit measures with reproducible results: The implicitMeasures package. *Journal of Open Source Software*, *5*(52), 2394. doi: 10.21105/joss.02394

Epifania, O. M., Robusto, E., & Anselmi, P. (2020, 2). Implicit social cognition through years: The Implicit Association Test at age 21.

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. doi: 10.1037/0022-3514.85.2.197

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. doi: 10.1037/0022-3514.93.3.353

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

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. doi: 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. doi: 10.1177/0146167204271418

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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. doi: 10.3102/10769986031002181

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209.