TPM®

RASCH GONE MIXED: A MIXED MODEL APPROACH TO THE IMPLICIT ASSOCIATION TEST

OTTAVIA M. EPIFANIA
EGIDIO ROBUSTO
PASQUALE ANSELMI
UNIVERSITY OF PADOVA

Despite the Implicit Association Test (IAT) is widely used for the implicit assessment of attitudes, the meaning of its effect remains unclear. Literature on the IAT has already highlighted the importance of the stimuli characteristics in influencing the meaning and the validity of the IAT measure. A model providing in-depth information at both respondents and stimuli levels can help in clarifying the meaning of the IAT measure. A modeling framework based on Linear Mixed Effects Models for a fine-grained analysis at both the respondent and the stimulus levels is presented. The proposed models provide a detailed picture of the contribution of each stimulus to the IAT effect, allowing for the identification of malfunctioning stimuli that can be eliminated or substituted to obtain better performing IATs. The information on respondents allows for a better interpretation of the IAT effect. Implications of the results and future research directions are discussed.

Keywords: Implicit social cognition; Implicit Association Test; Fully-crossed design; Rasch model; Log-normal model.

Correspondence concerning this article should be addressed to Ottavia M. Epifania, Department FISPPA, Section of Applied Psychology, University of Padova, Via Venezia 14, 35131 Padova (PD), Italy.
Email: marinaottavia.epifania@phd.unipd.it

The idea that people's attitudes include components of which they are aware (i.e., explicit or direct) and components of which they are not completely aware and that cannot be controlled (i.e., implicit or indirect) has now been widely accepted (e.g., Meissner, Grigutsch, Koranyi, Muller, & Rothermund, 2019). Among the measures aimed at capturing the indirect components of attitudes, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is one of the most studied and used in a constantly wider and more varied range of areas (for a review, see Epifania, Anselmi, & Robusto, 2021). By appropriately changing the labels of the attitude objects under investigation and leaving its structure unaltered, the IAT can be easily adapted for the investigation of different topics, ranging from personality and self-esteem (e.g., Van Tuijl et al., 2016; Vecchione et al., 2016) to emotions (Riediger, Wrzus, & Wagner, 2014), addiction behaviors (e.g., Tatnell, Loxton, Modecki, & Hamilton, 2019), and perception (e.g., Wu, Lu, van Dijk, Li, & Schnall, 2018). Given the IAT resistance to self-presentation strategies (Egloff & Schmukle, 2002; Greenwald, Poehlman, Uhlmann, & Banaji, 2009), its main applications are in social cognition, where it is used for the implicit assessment of attitudes toward different social groups (e.g., Anselmi, Vianello, & Robusto, 2011; Anselmi, Voci, Vianello, & Robusto, 2015), even in sensitive social contexts like hospitals (e.g., Zeidan et al., 2019). Despite its broad use, the meaning of the effect obtained from the IAT remains unclear. The aim of this contribution is to help in shedding light on the meaning of the IAT effect by considering the information that can be retrieved from stimuli and respondents' random variability.

TPM Vol. 28, No. 4, December 2021 – 467-483 – doi:10.4473/TPM28.4.5 – © 2021 Cises

The IAT is based on the speed and accuracy with which prototypical exemplars of two contrasting target categories (e.g., *White* and *Black* people in a Race IAT) and exemplars of two evaluative categories (*Good* and *Bad*) are sorted in the category to which they belong by means of two response keys. The categorization task takes place in two contrasting associative conditions. In one associative condition, the labels *Good* and *White* are displayed on the same side of the screen, and exemplars belonging to these categories are sorted with the same response key. The labels *Bad* and *Black* are displayed on the opposite side of the screen, and their exemplars are mapped with the same response key. In the contrasting associative condition, the labels *White* and *Black* switch their locations on the sides of the screen. *Good* and *Black* share the same side of the screen and are mapped with the same response key. *Bad* and *White* are displayed on the opposite side of the screen and are mapped with the other response key. The assumption underlying the IAT functioning is that respondents would show a better performance (i.e., faster response times and higher accuracy) when the task is consistent with their automatically activated association. The so-called *IAT effect* denotes the difference in respondents' performance between the two associative conditions.

The strength and direction of the IAT effect is usually expressed by the *D* score (Greenwald, Nosek, & Banaji, 2003), which results from the standardization of the difference in the average response time between the two conditions. The effect size measure proposed by Greenwald et al. (2003) is the most commonly used. Other authors have introduced modifications to the *D* score algorithm to either obtain more robust scores (Richetin, Costantini, Perugini, & Schönbrodt, 2015) or to fairly compare the IAT with other implicit measures (Epifania, Anselmi, & Robusto, 2020a). The *D* score provides general information on the implicit constructs that have been assessed, but it cannot inform about the automatic associations that mostly contribute to the IAT effect. Sticking with the Race IAT example, it would not be possible to discern whether the result is mostly due to an in-group favoritism, an out-group derogation, or even both. Moreover, since the *D* score is obtained by averaging across all trials in each associative condition, it cannot account for the dependency between the observations and the random variability due to both stimuli and respondents. As such, it might result in inflated scores (Brauer & Curtin, 2017; Wolsiefer, Westfall, & Judd, 2017), leading to inaccurate inferences on the implicit attitudes under investigation. Additionally, by overlooking the variability related to the stimuli, the information that can be gathered from each singular stimulus and their categories is completely neglected (Wolsiefer et al., 2017).

Different models have been proposed to get a better understanding of the IAT effect. Some of these models, like the Quad Model (Conrey, Gawronski, Sherman, Hugenberg, & Groom, 2005) or the ReAL Model (Meissner & Rothermund, 2013), consider only the accuracy responses, while other models, like the Diffusion Model (DM; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007) or the Discrimination-Association Model (DAM; Stefanutti, Robusto, Vianello, & Anselmi, 2013), simultaneously account for both accuracy and time responses. These models provide useful information at either the sample level (Quad model and ReAL model) or the respondent level (DAM and DM). DM and DAM also inform about the stimuli, but the information is provided at the level of stimuli categories and not at that of individual stimuli. Nevertheless, fine-grained information at the stimuli level would allow for testing whether individual stimuli are easily recognizable as prototypical exemplars of their own reference categories. Furthermore, the investigation on the contribution of each stimulus to the IAT effect would help in shedding light on the meaning of the implicit measure itself.

Rasch modeling (Rasch, 1960) of the IAT data can provide a fine-grained analysis at the level of each stimulus. Such an analysis allows for disentangling the automatic associations that mostly contribute to the IAT effect and provides a better understanding of the measure. For instance, by applying the Rasch model to the IAT discretized response times, Anselmi et al. (2011) found that positive words were those

that mostly contributed to the IAT effect. By analyzing responses to a Race IAT, the authors concluded that the implicit preference for European people over African people that is often observed in European respondents could be expression of ingroup favoritism rather than outgroup derogation. Despite the interesting insights provided by the Rasch modeling of IAT data, its application comes with some limitations. Firstly, the discretization of response times may result in a large loss of information. Additionally, the Rasch model is not able to account for the nonindependence of IAT observations, potentially resulting in biased parameter estimates and thus leading to an incorrect estimation of the importance of the effect of the IAT associative conditions (Judd, Westfall, & Kenny, 2017; McCullagh & Nelder, 1989). Finally, for the application of the Rasch model to the IAT, it was assumed that the difficulty of the two associative conditions did not differ across respondents, hence neglecting respondents' individual differences.

Linear Mixed-Effects Models (LMMs) can easily handle all the above-mentioned issues, while providing a Rasch parametrization of the data. LMMs also allow for treating the response times in their continuous nature, potentially avoiding the loss of information related to their discretization. To better understand the IAT effect and the meaning of the IAT measure while addressing the issues related to its sources of random variations, in the present work: (i) generalized LMMs (GLMMs) have been applied to IAT accuracy responses to obtain Rasch model parameter estimates; (ii) LMMs have been applied to IAT log-time responses to obtain log-normal model parameter estimates; and (iii) the relationship between the classic measure of the IAT effect (i.e., the *D* score) and the estimates of the model parameters obtained via the GLMM and the LMM has been investigated.

In the following section, the use of Rasch model and log-normal model for the analysis of IAT data is described, as well as the meaning of the resulting parameters. The application of these models to a Race IAT is presented. Some final remarks conclude the argumentation.

Models Specification

Accuracy and latency responses of the IAT can be modeled in a similar fashion by means of the Rasch model (Rasch, 1960) and the log-normal model (van der Linden, 2006), respectively.

In the Rasch model, the probability of a respondent to endorse the correct response (i.e., categorizing the stimulus into the correct category) can be expressed as a function of his/her ability θ (i.e., the ability to correctly categorize the stimuli) and stimuli easiness b (i.e., stimuli characteristics that make them more or less recognizable as prototypical exemplars of their category). The higher the value of θ, the higher the respondent's ability to perform the task, and, hence, the higher the proportion of stimuli correctly categorized. The higher the value of b, the easier the sorting of the stimulus in its own category. Thus, b informs about how much a stimulus is prototypical of the category that it is representing. Rasch model parameters estimates can be obtained by applying GLMMs to IAT accuracy responses. In GLMMs, the natural link function (g) between the linear combination of predictors and the observed values y is the *logit* (McCullagh & Nelder, 1989). The inverse of the link function g (i.e., g^{-1}) takes on a form that can be equated to the Rasch model (see De Boeck et al., 2011; Doran et al., 2007; Gelman & Hill, 2007 for the mathematical proofs).

The log-normal model allows for using the response times in their continuous nature by log-transforming the latencies. Consequently, the loss of information due to the discretization of the response times is avoided. According to this model, the log-time response of a respondent can be expressed as a function of respondent's speed τ (i.e., respondent's speed to categorize the stimuli) and stimuli time intensity δ (i.e., stimuli characteristics that make them require more or less time to get a response). The lower

the value of τ, the higher the respondents' speed. Likewise, the lower the value of δ, the lower the time the stimulus requires to get a response. As for the *b* parameters of the Rasch model, δ informs about how prototypical of its category the stimulus is. The lower the time it needs to be categorized, the more recognizable it is. Log-normal model parameter estimates can be obtained by applying LMMs to IAT response time after they have been log-transformed. In LMMs, the link between the predictors and the observed variables is the identity link, according to which the same scale of the dependent variable is taken as the scale for the link function, that is, the normal distribution.

The Best Linear Unbiased Predictors (BLUP) are used to obtain the Rasch model and log-normal estimates from the fitted (G)LMMs (De Boeck et al., 2011; Doran et al., 2007). BLUPs are the conditional modes of each level of the random effect, and they are not parameters of the model *per se*. They express the deviation of each level of the random effect from the estimated fixed effect. When added to the fixed effect of the IAT associative conditions, they result in the condition-specific estimates of either each respondent parameters or the condition-specific estimates of each stimulus parameters.

When using (G)LMMs to obtain the estimates of the Rasch model and the log-normal model parameters, the effect of the IAT condition on respondents' performance can be investigated by specifying the between–conditions and within-respondents variability, or, in other words, by specifying the random slopes of the respondents in the associative conditions. This results in condition-specific respondents' parameters. By specifying the between-conditions and within-stimuli variability (i.e., specifying the random slopes of the stimuli in the associative conditions), it is possible to obtain condition-specific estimates of the stimuli parameters, and hence investigate their contribution to the IAT effect. Three meaningful models for the analysis of the IAT accuracy responses were specified (left panel of Table 1), as well as three meaningful models for the analysis of the IAT log-time responses (right panel of Table 1). Besides the distribution of the error term, the GLMMs and the LMMs have the same random structures. The fixed intercept is set at 0, so that the fixed effects of the IAT associative conditions represent the expected average proportion of correct responses or the average response time in each condition for the Rasch model and the log-normal, respectively.

TABLE 1
Accuracy and log-time models overview

	Accuracy		Response time	
Model	Respondents	Stimuli	Respondents	Stimuli
1	Condition-specific ability (θ_{ik})	Overall easiness (b_j)	Condition-specific speed (τ_{ik})	Overall time intensity (δ_j)
2	Overall ability (θ_i)	Condition-specific easiness (b_{jk})	Overall speed (τ_i)	Condition-specific time intensity (δ_{jk})
3	Overall ability (θ_i)	Overall easiness (b_j)	Overall speed (τ_i)	Overall time intensity (δ_j)

Note. Respondent $i = 1,\ldots, I$, Stimulus $j = 1,\ldots, J$, Condition $k = 1,\ldots, K$, where I, J, and K, are the number of respondents, stimuli, and conditions, respectively.

The random structure specification of Model 1 (i.e., respondents' random slopes in the associative conditions and stimuli random intercept) results in the estimation of condition-specific respondents' parameters and overall stimuli parameters. The condition-specific respondents' parameters, either θ or τ, can

express if and how accuracy or speed performance of each respondent is affected by the IAT associative condition. By computing the difference between respondents' condition-specific parameters, a measure of the bias due to the associative conditions can be obtained, allowing for testing whether there is an effect of the condition on respondents' performance. Since the fixed intercept is set at 0 and stimuli are specified as random intercepts, their estimates are centered around 0, that is, the mean of the distribution of stimuli estimates.

The random structure specification of Model 2 (i.e., stimuli random slopes in the associative conditions and respondents' random intercept) results in the estimation of condition-specific stimuli parameters and overall respondents' parameters. This model allows for testing whether the functioning of the stimuli differs between conditions. If a stimulus shows a higher *b* (or δ) parameter in one condition than in the other, it means that it was easier (or required less time) to be categorized in the former condition rather than in the other. Moreover, the differential measure between the condition-specific stimuli parameters informs about the bias due to the associative conditions, hence providing information about the contribution of each stimulus to the IAT effect. Since the fixed intercept is set at 0 and respondents are specified as random intercepts, their estimates are centered around 0, that is, the mean of the distribution of respondents' estimates.

Finally, the random structure specification of Model 3 (i.e., stimuli random intercepts and respondents' random intercepts) results in the estimation of overall stimuli parameters and overall respondents' parameters. These parameters inform about the across-conditions performance of the respondents and the across-conditions functioning of the stimuli. This model should be preferred when a low between-conditions variability is observed at both respondents' and stimuli level. The lack of between-conditions variability already indicates that there is no IAT effect on either respondents' performance or stimuli characteristics. Since both respondents and stimuli are specified as random intercepts, their estimates are centered around 0. This model is not identified, at least for what concerns the Rasch model (see Gelman & Hill, 2007), and it is just used as a null model.

Response times must be log-transformed for the application of the log-normal model and for obtaining its estimates. From now on, the models applied on IAT accuracy responses will be identified with the letter "A," while the models applied on IAT log-time responses will be identified with the letter "T." The R code used for estimating these models is reported in the Appendix.

Outfit statistics were used to evaluate the fit of the data to the model chosen after model comparison. If *outfit* statistics ranged between 0.50 to 2.00 (Linacre, 2002), they express a good fit of the data to the model. However, the most problematic ones are the *outfit* statistics above 2, indicating a higher variability in the data that is not explained by the model (i.e., underfit). *Outfit* statistics below 0.50 indicates overfit of the model and will not be considered as problematic as those indicating underfit.

Method

The abovementioned models were applied to a Race IAT. Models were fitted with `lme4` package (Bates, Machler, Bolker, & Walker, 2015) in R (Version 3.5.1, R Core Team, 2018) and `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020b) was used for computing the IAT *D* score. A free and user-friendly tool for computing the IAT *D* score is retrievable at http://fisppa.psy.unipd.it/DscoreApp/ (Epifania, Anselmi, & Robusto, 2020c).

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%) self-identified as belonging to the Mediterranean ethnic group. A sensitivity power analysis was run with G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) to understand whether the sample size allowed ensuring 80% power to detect an effect size f^2 of at least 0.15 at $p < .05$. The sensitivity power analysis was run specifically for the investigation of the relationship between the parameter estimates of the Rasch and log-normal models and the IAT classic score and pointed out that the sample size was adequate for the aim.

Materials and Procedure

Participants were presented with a Race IAT. It was composed of 16 attribute stimuli, of which eight represented the *Good* category (i.e., "love," "good," "happiness," "joy," "glory," "peace," "pleasure," "laughter") and eight represented the *Bad* category (i.e., "bad," "pain," "failure," "annoying," "evil," "hate," "horrible," "terrible"). Target stimuli (same as in Study 2 by Nosek, Greenwald, & Banaji, 2005) were six faces of African people representing the *Black* category (three male and three female) and six faces of European people representing the *White* category (three male and three 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. The IAT administration included a built-in correction, for which participants had to correct each error response in order to go on with the experiment. They were instructed to be as accurate and fast as they could.

Data Cleaning and *D* score

Exclusion criteria based on both latency and accuracy responses were applied (Greenwald et al., 2003; Nosek, Banaji, & Greenwald, 2002). The algorithm *D*1 in Greenwald et al. (2003) was used for computing the *D* score. The difference was computed between the average response time in the BGWB and that in the WGBB condition: Positive scores stood for a possible preference for European people over African people. For the application of the LMMs to the log-time responses, the latencies at the incorrect responses were used.

Results

No participants or trials were eliminated grounding on the response time exclusion criteria. Three participants were excluded because of the accuracy deletion criterion (Nosek et al., 2002). The sample was finally composed of 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 in the WGBB condition was 667.11 ms (*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. When the latencies (expressed in sec-

ond) are log-transformed, the overall average response time is –0.29 log-seconds (*SD* = 0.40, skewness = 0.72, kurtosis = 3.88), the average response time in the WGBB condition is –0.43 log-seconds (*SD* = 0.31, skewness = 1.26, kurtosis = 3.73), and the average response time in the BGWB condition is –0.15 log-second (*SD* = 0.42, skewness = 0.24, kurtosis = 5.09). These response time distributions are consistent with computerized speed tasks like the IAT, where respondents are explicitly encouraged to give fast responses to all trials, and only a few numbers of slow responses are observed.

Rasch Models

Rasch models were obtained by applying GLMMs on IAT accuracy responses. Concerning Akaike Information Criterion (AIC), log-likelihood, and deviance, Model A2 (AIC = 3784.43, log-likelihood = –1886.21, deviance = 3722.43) performed better than Model A1 (AIC = 3786.51, log-likelihood = –1887.26, deviance = 3774.51) and Model A3 (AIC = 3785.87, log-likelihood = −1888.93, deviance = 3777.87). However, the latter one showed the lowest Bayesian Information Criterion (BIC) value (3813.53, 3825.91, 3828.00, BIC values for Model A3, A2, and A1, respectively). Model A2 was chosen. This model provided overall participants' ability parameters θ_i and condition-specific stimuli easiness parameters (b_{WGBB} and b_{BGWB}). Results from 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 that 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 ± 0.33). Seven respondents showed *outfit* statistics below 0.50, and they were retained in the analysis.

All stimuli showed appropriate *outfit* statistics in condition BGWB (*M* = 0.92 ± 0.12, Min = 0.69, Max = 1.08*). Outfit* statistics in condition WGBB (*M* = 0.94 ± 0.40, Min = 0.25, Max = 1.71*)* highlighted four stimuli with *outfit* statistics below 0.50, and they were retained in the analysis.

Stimuli easiness parameters for each condition resulting from Model A2 are reported in Table 2. The stimuli condition-specific easiness estimates are obtained by adding the condition-specific BLUP for each stimulus to the fixed effect of the associative condition.

The higher the value of *b*, the easier the stimulus is, meaning that it is easily recognized as belonging to its category and correctly assigned to that. Generally, IAT stimuli tended to be easy stimuli. Stimuli tended to be easier in the WGBB condition than in the BGWB condition, where they showed a higher easiness variability. On average, object stimuli in the WGBB condition were the easiest stimuli, while negative words stimuli tended to be the least easy stimuli in the BGWB condition, immediately followed by positive words in the same condition. The difference in stimuli easiness parameters is reported in Table 2 as well. Object stimuli showed the lowest average easiness difference, while attribute stimuli, particularly positive word stimuli, showed the highest average difference between conditions. The difference in the easiness estimates between the two associative conditions allowed for the identification of the stimuli of each category that gave the highest contribution and the least contribution to the IAT effect. The stimuli giving the highest contribution to the IAT effect were *joy* and *happiness* (*Good* category), *evil* and *horrible* (*Bad* category), *wm3* and *wf3* (*White* category), and *bm2* and *bf2 (Black* category). The stimuli giving the lowest contribution to the IAT effect were *love* and *glory* (*Good* category), *annoying* and *pain* (*Bad* category), *wf1* and *wm1* (*White* category) and *bm3* and *bf3* (*Black* category).

TABLE 2
Stimuli condition-specific estimates (b_{jk}) and overall time intensity estimates (δ_j)

	b_{WGBB}	b_{BGWB}	$b_{WGBB} - b_{WGBB}$	δ_j		b_{WGBB}	b_{BGWB}	$b_{WGBB} - b_{WGBB}$	δ_j
Positive words					*Negative* words				
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	3.45	1.83	1.62	0.03		3.28	1.79	1.49	0.04
SD	0.09	0.16	0.15	0.04		0.15	0.21	0.22	0.04
White faces					*Black* 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	3.54	2.35	1.19	–0.03		3.53	2.40	1.13	–0.07
SD	0.14	0.17	0.21	0.03		0.09	0.07	0.13	0.02

Note. b = easiness estimates obtained from Model A2; δ_j = time intensity estimates obtained from Model T3; wf = European female face; wm = European male face; bf = African female face; bm = African male face; WGBB = White-Good/Black-Bad condition; BGWB = Black-Good/White-Bad condition. Rows are ordered by decreasing values of $b_{WGBB} - b_{WGBB}$.

Log-Normal Models

Log-normal models were obtained by applying LMMs on IAT log-time responses. The three log-time models were compared between each other. Model T2 produced aberrant estimates (i.e., correlation between the stimuli random slopes equal to 1). Model T1 (AIC = 4399.66, BIC = 4448.06, log-likelihood = –2192.83, deviance = 4385.66) performed better than Model T3 (AIC = 4762.63, BIC = 4797.20, log-likelihood = −2376.32, deviance = 4752.63). Model T1 was chosen. This model resulted in 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 the associative conditions (M = 0.98 ± 0.01, Min = 0.98, Max = 0.99 for the BGWB condition, and M = 0.99 ± 0.01, *Min* = 0.98, *Max* = 1.03 for the WGBB condition). Concerning the stimuli, overall *Outfit* statistics indicated a good fit for all the stimuli (M =1.00 ± 0.16, Min = 0.77, Max =1.33). The condition-specific estimates of respondents' speed are obtained by adding the condition-specific BLUP of each respondent to the corresponding fixed effect of the associative conditions.

Responses in the WGBB condition tended to be 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 (σ^2 = 0.003), while the between-participants' variability was slightly higher in the BGWB condition (σ^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_j obtained from Model T3 are reported in Table 2. The stimuli time intensity estimates are obtained by adding each stimulus BLUP to the fixed intercept. Since the fixed intercept is set at 0, the time intensity estimates are centered around 0. The lower the value of δ_j, the lower the amount of time the stimulus needs to get a response. Attribute stimuli required more time to get a response, while object stimuli were the ones requiring less time, with exemplars of the *Black* category inducing the fastest responses. African American male faces required less time to obtain a response than African American female faces did, while this pattern was not observed for White American people faces. 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 words (*failure*, *annoying*, *pain*) showed higher time intensity estimates than the other negative words. Object stimuli tended to have similar time intensity estimates.

Regression Model: *D* score

A *speed-differential* measure was computed by taking the difference between speed estimates in the BGWB condition and speed estimates in the WGBB condition. Negative values indicated a respondent faster 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 ($r = .13$, $p = .32$), while it poorly and negatively correlated with the *speed-differential* ($r = –.14$, $p = .28$). Ability moderately correlated with the speed parameters in the WGBB condition ($r = .32$, $p = .01$).

Participants' ability and *speed-differential* were regressed on the *D* score. Backward deletion was used to investigate the linear combination of predictors accounting for the higher proportion of explained variance. Backward deletion kept both the predictors in the model, which accounted for about 80% of the total variance — Adjusted $R^2 = .78$, $F(2, 59) = 106.30$, $p < .001$. *Speed-differential* strongly and positively predicted *D* score — $B = 1.93$, $t(59) = 13.88$, $p < .001$. Ability negatively predicted the *D* score — $B = –0.18$, $t(59) = –2.48$, $p = .012$.

To better understand the specific contribution of the speed of each associative condition, a model including the linear combination of ability estimate, speed estimate in the WGBB condition, and speed estimate in the BGWB condition was specified as well. Backward deletion kept all three predictors in the model, which accounted for almost 80% of the total variance — Adjusted $R^2 = .79$, $F(3, 58) = 76.46$, $p < .001$. Speed estimate in the WGBB condition negatively predicted the *D* score, $B = –2.22$, $t(58) = –11.43$, $p < .001$, while speed in the BGWB condition positively predicted it — $B = 1.92$, $t(58) = 14.16$, $p < .001$. Despite the ability parameter remained in the model, its contribution was no longer significant — $B = –0.13$, $t(58) = –1.76$, $p = .08$.

Final Remarks

The application of (G)LMMs to IAT data proved to be an effective modeling framework for obtaining the estimates of Rasch model and log-normal model parameters while accounting for the nonindependence of the observations.

The fine-grained analysis at the stimuli level allowed a deeper understanding of the meaning of the IAT measure, for example by giving the chance to investigate the stimuli that were not representative of their category or did not contribute to the IAT effect. Specifically, these models provided detailed information about how much each stimulus is representative of its own category. According to Nosek et al. (2005), a valid IAT measure can be obtained by using as few as two stimuli to represent each category. The information at the stimuli level provided by these models allows for exploiting the most representative and prototypical exemplars of each category. For instance, it was possible to identify two stimuli for each category providing the highest information (e.g., the words *joy* and *happiness* for the *Good* category). Grounding on these results, it is possible to design new IATs that can maximize the information, while reducing the number of stimuli representing each category and, consequently, the number of trials. However, the estimates provided by the Rasch model and the log-normal model were not considered together, and hence the information they are providing should be interpreted with caution. This issue can be addressed by using a hierarchical approach like the one in van der Linden (2006).

The representativeness of the stimuli can be pretested in a sample drawn from the population of interest. Even though this procedure is a valid procedure, it should be repeated every time the IAT is used on samples drawn from different populations. One of the advantages of Rasch modeling is that the estimates obtained on the stimuli are independent from the sample from which they were estimated. As such, stimuli parameter estimates can provide information on stimuli functioning that can be generalized to other samples (drawn from the same population) than the one from which they were obtained. Besides, by using this approach, it is possible to add new stimuli and test their functioning independently from the functioning of the old stimuli.

The information at the stimuli level can also be used for understanding the associations mostly driving the IAT effect. In this case, the evaluative dimensions *Good* and *Bad* were the stimuli categories showing the highest difference between the associative conditions. Both stimuli categories resulted easier in the WGBB condition than in the BGWB condition, meaning that the *Good* stimuli were more easily sorted when their category shared the response key with *White* category than when it shared the response key with *Black* category. Similarly, *Bad* stimuli were more easily sorted when their category shared the response key with *Black* category than when it shared the response key with *White* category. This result is in line with the positive primacy effect found by Anselmi et al. (2011), and it also highlights the contribution of the negative evaluative dimension in influencing the IAT effect. Given that the IAT effect appears to be mostly driven by evaluative dimensions, this result is in contrast with what was found by Klauer et al. (2007), according to whom attitudes influence the performance at the IAT through the categorization of the object stimuli.

These models also resulted in detailed information on respondents' accuracy and speed performance. Understating how respondents are behaving during the IAT administration is crucial to get a deeper comprehension of its measure and on the factors that might influence it. Respondents' accuracy performance was not affected by the IAT associative conditions, while their speed performance was. Consequently, the IAT effect seems to be mostly due to a respondents' slowdown, while the accuracy performance remains unaltered. This result can be interpreted by considering the speed-accuracy trade-off (Klauer et al., 2007). Indeed, respondents tend to slow down to maintain the accuracy unaltered in the condition that is against their automatically activated associations.

Not surprisingly, the *D* score was strongly related with the speed parameters, both *speed-differential* and condition-specific speed estimates, while the contribution of ability was negligible. By using a differential measure to predict the *D* score, it is not possible to understand the actual weight of each associative condition in determining the final score. Conversely, when the condition-specific estimates were used to predict the *D*

score, it was possible to isolate and highlight the higher contribution of the speed estimate pertaining to the WGBB condition compared with those pertaining to BGWB condition. This result is consistent with those obtained from the stimuli easiness estimates.

Given their flexibility, these models can be used for modeling data from other implicit measures similar to the IAT, such as the Single Category IAT (SC-IAT; Karpinski & Steinman, 2006) or the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001). Since the SC-IAT results from a slight modification of the IAT procedure and is based on speed and accuracy of stimuli categorization, both the accuracy and the log-normal models can be used for modeling its responses. Differently, the GNAT is based solely on accuracy responses. Given that the accuracy and the log-time models do not rely on each other to be applied, it is possible to use only the accuracy models for obtaining the estimates of the Rasch model parameters on the GNAT accuracy responses. Moreover, since the IAT can be used together with either the SC-IAT (e.g., Karpinski & Steinman, 2006; Chevance, Stephan, Heraud, & Boiché, 2018) or the GNAT (e.g., Ueda, Yanagisawa, Ashida, & Abe, 2017; Yang, Zhao, Guan, & Huang, 2017), it is possible to specify LMMs able to simultaneously account for the different implicit measures in one comprehensive model.

Since the aim of the study was to investigate the effect of the IAT associative condition on respondents' performance or stimuli functioning within a Rasch approach, no other predictors were entered in the models. However, given the flexibility of these models, it is possible to include other fixed effects for the investigation of the effect of different features of the stimuli (e.g., whether it is a word or an image) or of different characteristics of the respondents.

In this study, we did not investigate and compare the relationship between explicit measures of attitudes, behavioral outcomes, estimates obtained through Rasch and log-normal models, and *D* score. It can be speculated that, since the estimates obtained from the (G)LMMs are not influenced by unwanted error variance due to the nonindependence of the observations, they can be more reliable than the *D* score, hence allowing for a better inference of the construct under investigation. Therefore, they may result in a better prediction of behavioral outcomes, as well as showing stronger relations with explicit evaluations tapping the same construct. Future studies should address this issue.

Rasch analysis based on small samples, such as that used in this study, should be used for exploratory purposes with extreme caution (Chen, Lederking, Jin, Wyrwich, Gelhorn & Revicki, 2014). Nonetheless, when LMMs are employed, it is not the sample size *per se* that matters, but the number of observations for each unit of analysis, in this case, the respondents. There were 120 observations for each respondent, which should have ensured reliable estimates for the respondents.

This work highlighted how a simple approach can lead to a thorough and detailed analysis of the IAT data within a Rasch framework. The fine-grained analysis at the stimulus, the participant, and the associative condition levels provided by these models may lead to new interesting insights on the IAT functioning and meaning.

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. https://doi.org/10.1027/1618-3169/a000106

Anselmi, P., Voci, A., Vianello, M., & Robusto, E. (2015). Implicit and explicit sexual attitudes across genders and sexual orientations. *Journal of Bisexuality*, *15*(1), 40-56. https://doi.org/10.1080/15299716.2014.986597

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

Brauer, M., & Curtin, J. J. (2017). Linear mixed-effects models and the analysis of non-independent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389-411. https://doi.org/10.1037/met0000159

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485-493. https://doi.org/10.1007/s11136-013-0487-5

Chevance, G., Stephan, Y., Heraud, N., & Boiché, J. (2018). Interaction between self-regulation, intentions and implicit attitudes in the prediction of physical activity among persons with obesity. *Health Psychology*, *37*(3), 257. https://doi.org/10.1016/j.psychsport.2017.04.007

Conrey, F., Gawronski, B., Sherman, J., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469-487. https://doi.org/10.1037/0022-3514.89.4.469

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, *39*(12), 1-28. https://doi.org/10.18637/jss.v039.i12

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/10.1111/j.1467-9868.2007.00600.x

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology*, *83*(6), 1441. https://doi.org/10.1037//0022-3514.83.6.1441

Epifania, O. M., Anselmi, P., & Robusto, E. (2020a). A fairer comparison between the Implicit Association Test and the Single Category –Implicit Association Test. *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, *27*(2), 207-220. https://doi.org/10.4473/TPM27.2.4

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

Epifania, O. M., Anselmi, P., & Robusto, E. (2020c). DscoreApp: A Shiny Web Application for the Computation of the Implicit Association Test *D* Score. *Frontiers in Psychology*, *10*, 2938. https://doi.org/10.3389/fpsyg.2019.02938

Epifania, O. M., Anselmi, P., & Robusto, E. (2021). Implicit social cognition through the years: The Implicit Association Test at age 21. *Psychology of Consciousness: Theory, Research, and Practice.* Advance online publication. https://doi.org/10.1037/cns0000305

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge: Cambridge University Press.

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

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

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

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

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16-32. https://doi.org/10.1037/0022-3514.91.1.16

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

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

McCullagh, P., & Nelder, J. (1989). *Generalized linear models.* (2nd ed.). New York, NY: Chapman & Hall.

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

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of Associations and Recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*(1), 45-69. https://doi.org/10.1037/a0030734

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625-666. https://doi.org/10.3758/BRM.42.4.944

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

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

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

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling Implicit Association Test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PloS One*, *10*(6). https://doi.org/10.1371/journal.pone.0129601

Riediger, M., Wrzus, C., & Wagner, G. (2014). Happiness is pleasant, or is it? Implicit representations of affect valence are associated with contrahedonic motivation and mixed affect in daily life. *Emotion*, *14*(5), 950-961. https://doi.org/10.1037/a0037711

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

Tatnell, D., Loxton, N., Modecki, K., & Hamilton, K. (2019). Testing a model of reward sensitivity, implicit and explicit drinker identity and hazardous drinking. *Psychology and Health*, *34*(12):1407-1420. https://doi.org/10.1080/08870446.2019.1606221

Ueda, R., Yanagisawa, K., Ashida, H., & Abe, N. (2017). Implicit attitudes and executive control interact to regulate interest in extra-pair relationships. *Cognitive, Affective and Behavioral Neuroscience*, *17*(6), 1210-1220. https://doi.org/10.3758/s13415-017-0543-7

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

Van Tuijl, L., Glashouwer, K., Bockting, C., Tendeiro, J., Penninx, B., & De Jong, P. (2016). Implicit and explicit self-esteem in current, remitted, recovered, and comorbid depression and anxiety disorders: The NESDA study. *PloS One*, *11*(11), e0166116. https://doi.org/10.1371/journal.pone.0166116

Vecchione, M., Dentale, F., Alessandri, G., Imbesi, M., Barbaranelli, C., & Schnabel, K. (2016). On the applicability of the Big Five Implicit Association Test in organizational settings. *Current Psychology*, *36*, 665-674. https://doi.org/10.1007/s12144-016-9455-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. https://doi.org/10.3758/s13428-016-0779-0

Wu, Y., Lu, J., van Dijk, E., Li, H., & Schnall, S. (2018). The color red is implicitly associated with social status in the United Kingdom and China. *Frontiers in Psychology*, *9*(OCT). https://doi.org/10.3389/fpsyg.2018.01902

Yang, Q., Zhao, Y., Guan, L., & Huang, X. (2017). Implicit attitudes toward the self over time in Chinese undergraduates. *Frontiers in Psychology*, *8*(OCT). https://doi.org/10.3389/fpsyg.2017.01914

Zeidan, A., Khatri, U., Aysola, J., Shofer, F., Mamtani, M., Scott, K., Conlon, L. W., & Lopez, B. (2019). Implicit bias education and emergency medicine training: Step one? Awareness. *AEM Education and Training*, *3*(1), 81-85. https://doi.org/10.1002/aet2.10124

APPENDIX A

The Rasch and log-normal estimates were obtained by means of lme4 package (Bates et al., 2015) in R. The R code used for estimating the models and for extracting the parameter estimates is illustrated in this Appendix. This code can be copied and pasted in an R script, and it can be executed without changes as long as the data set on which the models are applied has the following characteristics:

- `subject`: Column containing the respondents' IDs (can be numeric, a factor, or a string, as long as it is unique for each respondent).
- `condition`: Column containing the labels for the two associative conditions of the IAT (factor with two levels such as `mappingA` and `mappingB`).
- `stimuli`: column containing the labels identifying each stimulus (e.g., `good`, `bad`, `wf1`, `bm2`).
- `latency`: Column containing the latency of the IAT responses. Latency can be expressed in seconds or milliseconds (in this paper, we used seconds). In case the IAT included a built-in correction for the error responses, the raw response times should be used instead of the corrected ones.
- `correct`: Column containing the accuracy of the IAT responses, where `0` is the incorrect response and `1` is the correct response.

The data set must be in a long format. This means that the response of each respondent on each stimulus in each associative condition must be on a separate row, and the total number of observations (and rows) for each subject must correspond to the total number of critical trials in the two associative conditions. For instance, in this study participants were presented with 60 trials in each associative condition, so that we had 120 trials for each respondent, and consequently 120 rows for each participant.

In both accuracy and log-time responses, the fixed intercept was set at 0, so that the estimates for the effect of the IAT associative conditions can be interpreted as the expected log-odds of the probability of a correct response in each condition or the expected average log-response time in each condition, respectively. For both accuracy and log-time responses, in Model 1 (Table 1) the estimates of the stimuli are centered at 0 (argument `(1|stimuli)`), while in Model 2 (Table 1) respondents' estimates are centered at 0 (argument `(1|subject)`).

Accuracy Models Specification

The code for the specification of the accuracy models is illustrated. The name of the data set in the argument `data` must be changed accordingly.

Model 1: Between-stimuli variability specified as random intercepts (i.e., `(1|stimuli)`). Within-subjects and between-conditions variability specified as random slopes of the respondents in the conditions (i.e., `(0 + condition|subject)`).

```
library(lme4) # upload the package for the estimation of the mod-
els
a1 <- glmer(correct ~ 0 + condition + (1|stimuli) +
          (0 + condition|subject),
```

```
            data = your_data, family = "binomial")
summary(a1) # summary of the results
```

Model 2: Between-subjects variability specified as random intercepts (i.e., `(1|subject)`). Within-stimuli and between-conditions variability specified as random slopes of the stimuli in the conditions (i.e., `(0 + condition|stimuli)`).

```
a2 <- glmer(correct ~ 0 + condition + (1|subject) +
           (0 + condition|stimuli),
            data = your_data, family = "binomial")
summary(a2) # summary of the results
```

Model 3: Between-subjects variability specified as random intercepts (i.e., `(1|subjects)`). Between-stimuli variability specified as random intercepts (i.e., `(1|stimuli)`).

```
a3 <- glmer(correct ~ 0 + condition + (1|stimuli) + (1|subject),
            data = your_data,
            family = "binomial")
summary(a3) # summary of the results
```

Once the three models have been estimated, they can be compared with each other. Model 1 (`a1`) and Model 2 (`a2`) have the same degrees of freedom:

```
anova(a1, a2, a3)
```

Accuracy Models: Rasch Model Parameter Estimates

Grounding on the results of the model comparison, the best fitting model can be selected for extracting the Rasch model parameter estimates.

Model 1 results in condition-specific respondents' estimates and overall stimuli estimates. Respondents' condition-specific ability estimates can be extracted as follows:

```
cond_ability <- coef(a1)$subject[, -1] # drop the first column
                        # (fixed intercepts set at 0)
                        # rownames are the subjects' IDs
```

Stimuli easiness estimates can be extracted and stored in a data frame as well:

```
easiness <- data.frame(
              stimuli = rownames(coef(a1)$stimuli),
```

```
                easiness = coef(a1)$stimuli[, 1] # select only the
                        # random estimates intercept
)
```

Model 2 results in condition-specific stimuli estimates and overall respondents' estimates. Stimuli condition-specific estimates can be extracted as follows:

```
easiness_cond <- coef(a2)$stimuli[, -1] # drop the first column
                        # (fixed intercept set at 0)
                        # rownames are stimuli labels
```

Respondents overall ability estimates can be extracted and stored in a data frame:

```
ability <- data.frame(
             subject = rownames(coef(a2)$subject),
             ability = coef(a2)$subject[, 1] # select only the
             # random intercept estimates
)
```

Model 3 results in overall respondents' estimates and overall stimuli parameters. Respondents overall ability estimates can be extracted and stored in a data frame:

```
ability <- data.frame(
             subject = rownames(coef(a3)$subject),
             ability = coef(a3)$subject[, -1]
)
```

Stimuli overall easiness estimates can be extracted and stored as well:

```
easiness <- data.frame(
              stimuli = rownames(coef(a3)$stimuli),
              easiness = coef(a3)$stimuli[, -1]
)
```

Log-Time Models Specification

The code for the estimation of the log-normal models is the same as the one used for estimating the Rasch models. The changes concern the name of the specific function to use (from `glmer()` to `lmer()`) and the dependent variable (from `correct` to `log(latency)`). For this reason, we report the code for the estimation of Model 1 only.

```
t1 <- lmer(log(seconds) ~ 0 + condition + (1|stimuli) +
          (0 + condition|subject),
```

```
            data = your_data,
            REML = FALSE) # Maximum Likelihood estimation
summary(t1) # summary of the results
```

For log-time models comparison, the same code as the one used for accuracy models comparison can be used by changing the names of the models from `a` to `t`.

Log-Time Models: Log-Normal Model Parameters

We report the code for extracting the log-normal model estimates for log-time Model 1, assuming it was the best fitting model according to model comparison. The same code used for extracting the estimates for the accuracy models can be used for extracting the parameters of the log-normal models. The changes regard the name of the objects containing the models, from `a` to `t`, and the names of the new objects created for the parameters (e.g., from `easiness` to `intensity`).
Respondents' condition-specific parameters:

```
cond_speed <- coef(a1)$subject[, -1] # drop the first column
                      # (fixed intercepts set at 0)
                      # rownames are the subjects' IDs
```

Stimuli overall time intensity parameters:

```
intensity <- data.frame(
               stimuli = rownames(coef(t1)$stimuli),
               intensity = coef(t1)$stimuli[, 1] # select only
the
            # random intercept estimates
)
```