Rasch gone mixed: A mixed model approach to the Implicit Association Test

Ottavia M. Epifania, Egidio Robusto, and Pasquale Anselmi

Department of Philosophy, Sociology, Education, and Applied Psychology,
University of Padova

Author Note

Ottavia M. Epifania <https://orcid.org/0000-0001-8552-568X>

Egidio Robusto <https://orcid.org/0000-0002-7583-2587>

Pasquale Anselmi <https://orcid.org/0000-0003-2982-7178>

The Authors would like to thank Dr. Paul DeBoeck for his suggestions on a previous version of the paper.

The authors have no conflict of interest to disclose.

Correspondence concerning this Article should be addressed to: Ottavia M. Epifania, Department of Philosophy, Sociology, Education, and Applied Psychology, Via Venezia 14, Padova, Italy. E-Mail: marinaottavia.epifania@phd.unipd.it

Rasch gone mixed: A mixed model approach to the Implicit Association Test

Abstract

The Implicit Association Test (IAT) is one of the most commonly used measures for the implicit assessment of preferences and attitudes. Nevertheless, the meaning of the IAT effect remains unclear. Since the IAT is based on the speed and accuracy with which different stimuli are categorized in the category to which they belong, the sense of the IAT measure is strongly related to the functioning of the stimuli used to represent the categories. Thus, a model providing a fine-grained analysis at both respondents and stimuli levels would help in clarifying and interpreting the IAT measure. By applying Linear Mixed Effects Models to IAT accuracy and time responses, this study aims to present a modeling framework for IAT data resulting in such a detailed analysis at both respondents and stimuli levels. 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. Additionally, the detailed information on respondents' performance yields a better understanding of the classic measure of the IAT effect (i.e., the *D* score). Implications of the results and future research directions, also implying different implicit procedures, are discussed.

Keywords: Implicit Social Cognition; Implicit Association Test; Rasch Model; Log-normal Model

Rasch gone mixed: A mixed model approach to the Implicit Association Test

Introduction

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, Müller, & Rothermund, 2019). Among the measures aimed at capturing the implicit 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 fields of application (see Epifania, Robusto, & Anselmi, 2020, for a review). 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), it finds its main application in social cognition, where it is used for the implicit assessment of attitudes towards 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. For example, in an IAT for the investigation of the implicit attitudes towards Black and White people (i.e., Race IAT), the implicit preference for White people over Black people could be due to either a favoritism for White people, or a derogation for Black people, or both. The aim of this contribution is to help in shedding a light on the meaning of

the IAT effect by considering the information that can be retrieved from the random variability at the levels of the stimuli and the respondents.

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, under the assumption that respondents would have a better performance, in terms of faster response times and higher accuracy, when the task is compatible with their automatically activated association. 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 so-called *IAT effect* results from the difference in respondents' performance between the two 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, 2020b). The *D* score provides general information on the implicit constructs under investigation, but it cannot inform about the automatic associa-

tions 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 single IAT observations due to the random variability at the levels of both the stimuli and the respondents. As such, it might result in inflated scores (Brauer & Curtin, 2017; Judd, Westfall, & Kenny, 2012; 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 for getting 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. The functioning of the stimuli is indeed a vital component for the correct functioning of the IAT itself (e.e., Bluemke & Friese, 2006). As such, fine-grained information at the stimuli level would allow for testing the functioning of the individual stimuli, resulting in better functioning IATs. Furthermore, the investigation on the contribution of each stimulus to the IAT effect would help in shedding a light on the meaning of the implicit measure itself.

The Rasch modeling (Rasch, 1960) of the IAT data provides such a fine-grained analysis at the level of each stimulus. It allows for disentangling the automatic associations that mostly contribute to the IAT effect and for providing a better understanding of the measure itself. For instance, by applying the Rasch model to the IAT discretized response times, Anselmi et al. (2011) found that positive words were the stimuli that mostly contributed to the IAT effect. By drawing on this result, the authors suggested that the implicit preference for White people over Black people that is often observed in White respondents could be expression of in-group favoritism rather than out-group derogation. Despite the interesting insights provided by the Rasch modeling of the IAT data, its application comes with some limitations. Firstly, the discretization of the response times may result in a large loss of information that can be avoided by considering the response times in their continuous nature. Additionally, the Rasch model in its typical form cannot account for the non-independence of the IAT observations. This potentially results in biased parameters estimates and might lead to an incorrect estimation of the importance of the effect of the IAT associative conditions (Judd et al., 2012; 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 the respondents' individual differences were neglected.

Linear Mixed Effects Models (LMMs) can easily handle all the above-mentioned issues, while providing a Rasch parametrization of the data. LMMs also allows 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 parameters estimates; (ii) LMMs have been applied to IAT log-time responses to obtain

log-normal model parameters estimates; and (iii) The relationship between the classic measure of the IAT effect (i.e., the D score) and the estimates of the models 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

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

In the Rasch model, the probability of observing a correct response (i.e., the stimulus is sorted into the correct category) can be expressed as a function of the respondent's ability θ (i.e., the ability of the respondent to correctly categorize the stimuli) and the stimulus easiness b (i.e., the characteristics of the stimulus that make it more recognizable as a prototypical exemplar of its own 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 stimulus is, or, in other words, it is easily sorted in the category to which it belongs and hence can be considered as a prototypical exemplar of that category. The interplay between these two parameters determine the probability of a correct response. The estimates of the Rasch model parameters 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, Bates, Bliese, & Dowling, 2007; Gelman & Hill, 2007, for the mathematical proofs).

The log-normal model (van der Linden, 2006) allows for using the response time in their continuous nature by considering the normal distribution of the log-time responses. 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 the respondent's speed τ (i.e., the speed of the respondent to categorize the stimuli) and the stimulus time intensity δ (i.e., the characteristics of the stimulus that make it require less time for getting 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 for getting a response. The time intensity parameter δ informs the representativeness of the stimulus of its own category. The lower the time it needs to be categorized, the more representative of its own category it is. The interplay between these two parameters determine the log-time responses. The estimates of the log-normal model parameters can be obtained by applying LMMs to the IAT log-response time. 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 for obtaining 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.

In (G)LMMs, the effect of the IAT associative 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., the random slopes of the stimuli in the associative conditions), it is possible to obtain condition-specific estimates of the stimuli parameters. This allows for investigating the contribution of the stimuli 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). The random structure specified for the GLMMs and the LMMs were identical. The difference between the specification of the GLMMs and the LMMs lies in the assumption made on the distribution of the error term. In the former case, it is assumed to follow a logistic distribution (i.e., $\varepsilon_i \sim \mathcal{L}(0, \sigma_i^2)$, where $\mathcal{L}$ denotes the logistic distribution as in Doran et al. (2007)), in the latter one it is assumed to follow a normal distribution (i.e., $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$). The fixed intercept is set at 0, so that the estimated fixed effects of the IAT associative conditions represent the expected *log-odds* of a correct response or the average log- response time in each associative condition for the Rasch model and the log-normal, respectively.

Model 1 is considered as the Null Model. The random structure specification of Model 1 (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

Table 1: Rasch model and log-normal model estimates.

Model	Rasch model		Log-normal 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: Respondent $p \in \{1, \dots, P\}$, Stimulus $s \in \{1, \dots, S\}$, Condition $c \in \{1, \dots, C\}$, where P , S , and C , are the number of respondents, stimuli, and conditions, respectively.

estimates are centered around 0.

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 results as the best fitting one when a high within-stimuli between-conditions variability is observed, along with a low within-respondents between-conditions variability. This model allows for testing whether the functioning of the stimuli differs between conditions. If a stimulus shows a higher b_s (or δ_s) 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 estimates informs about the bias due to the associative conditions, which in turn provides 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.

The random structure specification of Model 3 (i.e., respondents' random slopes in the asso-

ciative conditions and stimuli random intercept) results in the estimation of condition-specific respondents' parameters and overall stimuli parameters. This model results as the best fitting one when a high within-respondents between-conditions variability is observed, along with a low within-stimuli between-conditions variability. The estimates of the condition-specific respondents' parameters, either θ_p or τ_p , express if and how accuracy or speed performance of each respondent was affected by the IAT associative condition. By computing the difference between respondents' condition-specific estimates, 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 the stimuli estimates.

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 .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 (i.e., the data shows less variability than that expected by the model) and will not be considered as problematic as those indicating underfit.

Method

The above-mentioned models were applied to a Race IAT. Is were fitted with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R (Version 3.5.1, R Core Team, 2018) and the `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020c) 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, 2020a).

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. A sensitivity power analysis was run with GPower (Faul, Erdfelder, Buchner, & Lang, 2009) to understand whether the sample size allows for 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 models parameters estimates 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 by 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 by Nosek, Greenwald, & Banaji, 2005) were 6 faces of Black people (3 male and 3 female) and 6 faces of White People (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. 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 *DI* 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 the WGBB condition: Positive scores stood for a possible preference for White people over Black people. In applying the LMMs, 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 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 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 second) were log-transformed, the overall average response time was -0.29 log-seconds ($sd = 0.40$, $skewness = 0.72$, $kurtosis = 3.88$), the average response time in the WGBB condition was -0.43 log-seconds ($sd = 0.31$, $skewness = 1.26$, $kurtosis = 3.73$), and the average response time in the BGWB condition was -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 en-

couraged to give fast responses to all trials, and only a few numbers of slow responses are observed.

Rasch models.

The estimates of the Rasch model parameters were obtained by applying GLMMs on IAT accuracy responses. 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}). Results from Model A2 indicated a higher probability of correct response in the WGBB condition ($\log\text{-odds} = 3.45$, $SE = 0.12$) than in the BGWB condition ($\log\text{-odds} = 2.07$, $SE = 0.11$)¹. 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 ($r = .34$).

The 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 2.

¹ $\log\text{-odds}$ are the effect size of the probability for a correct response.

Table 2: Stimuli condition-specific easiness parameters (b_{sc}) and overall time intensity parameters (δ_s) - Race IAT

	b_{WGBB}	b_{BGWB}	$b_{WGBB} - b_{WGBB}$	δ_s		b_{WGBB}	b_{BGWB}	$b_{WGBB} - b_{WGBB}$	δ_s
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
<i>M (SD)</i>	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					Black people				
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
<i>M (SD)</i>	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 female face, “wm”: White male face, “bf”: Black female face, “bm”: Black male face; WGBB: White-Good/Black-Bad condition; BGWB: Black-Good/White-Bad condition. Rows are ordered by decreasing values of $b_{WGBB} - b_{WGBB}$. The units of parameter b are the log-odds. The units of parameter δ are the log-seconds.

Overall, 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 estimates is reported in Table 2 as well. Object stimuli showed the lowest average easiness difference, while attribute stimuli, particularly *Good* exemplars, 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 or the least contribution to the IAT effect. The stimuli giving 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 giving 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.

Log-normal models were obtained by applying GLMMs 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 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 resulted in condition-specific participants' speed parameters (τ_{WGBB} and τ_{BGWB}) and overall stimuli time intensity parameters δ_s . 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 tended to be faster ($\beta = -0.43$, $SE = 0.02$) than responses in the BGWB condition ($\beta = -0.15$, $SE = 0.03$). The between-stimuli variability was particularly low ($\sigma = 0.003$), while the between-participants variability was slightly higher in the BGWB condition ($\sigma = 0.05$) than that in the WGBB one ($\sigma = 0.02$). The correlation between respondents' variability in the two conditions was strong ($r = .63$).

Stimuli time intensity parameters δ_s 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_s , the lower the amount of time the stimulus needs for getting a response. Attribute stimuli required more time to get a response than object stimuli. *Black* male faces required less time for getting a response than *Black* female faces. This pattern was not observed for *White* 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 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 *D* score

A *speed-differential* measure was computed as the difference between speed parameters in the BGWB condition and 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*. Partici-

pants' ability poorly and positively correlated with speed in the BGWB condition ($r = .13$, $p = .32$), and it correlated negatively and poorly with the *speed-differential* ($r = -.14$, $p = .28$), although these correlations were not significant. Ability moderately correlated with the speed estimate 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 70% of the total variance ($Adjusted R^2 = .78$, $F(2, 59) = 106.3$, $p < .001$). *Speed-differential* strongly and positively predicted D score ($\beta = 1.93$, $t(59) = 13.88$, $p < .001$). Ability negatively predicted the D score ($\beta = -0.18$, $t(59) = -2.48$, $p = .012$).

To better understand the specific contribution of the speed of each associative condition, a model including the linear combination of 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 ($B = -0.13$, $t(58) = -1.76$, $p = .08$).

Final remarks

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

The fine-grained analysis at the stimuli level allowed for a deeper understanding of the meaning of the IAT measure, for example by giving the chance of investigating 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 category *Good*). 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. As such, 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 (2007).

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 parameters 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 has been 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 for getting 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 negli-

gible. 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). The SC-IAT results from a slight modification of the IAT procedure and is based on speed and accuracy of stimuli categorization. Consequently, both the accuracy and the log-normal models can be used for modeling its responses. Differently, the GNAT is solely based 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, the IAT can be used together with either the SC-IAT (Epifania, Anselmi, & Robusto, 2020b; Karpinski & Steinman, 2006). This modeling approach allows for extending the random structure if the (G)LMMs to include multiple implicit measures into 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 non-independence 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 et al., 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 stimuli, participants, and 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. doi: 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. doi: 10.1080/15299716.2014.986597
- 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
- Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, 42(2), 163–176. doi: 10.1016/j.jesp.2005.03.004
- Brauer, M., & Curtin, J. J. (2017). Linear mixed-effects models and the analysis of nonindependent 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. doi: 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. doi: 10.1007/s11136-013-0487-5
- 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. doi: 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.
- 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. doi:

10.1111/j.1467-9868.2007.00600.x

Epifanio, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of personality and social psychology*, 83(6), 1441. doi: 10.1037//0022-3514.83.6.1441

Epifania, O. M., Anselmi, P., & Robusto, E. (2020a). Dscoreapp: A shiny web application for the computation of the implicit association test d-score. *Frontiers in Psychology*, 10, 2938. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2019.02938> doi: 10.3389/fpsyg.2019.02938

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

Epifania, O. M., Anselmi, P., & Robusto, E. (2020c). implicitMeasures: Computes the Scores for Different Implicit Measures [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=implicitMeasures> (R package version 0.2.0)

Epifania, O. M., Robusto, E., & Anselmi, P. (2020, 2). Implicit social cognition through years: The Implicit Association Test at age 21. Retrieved from <https://advance.sagepub.com/articles/Implicit-social-cognition-through> doi: 10.31124/advance.11914416.v1

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. doi: 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. doi: 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. doi: 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. doi: 10.1037/a0015575
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69. doi: 10.1037/a0028347
- 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. doi: 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. doi: 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. 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.
- McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. (2nd ed.). New York: Chapman & Hall.
- Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. *Frontiers in Psychology*, 10. doi: 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. doi: 10.1037/a0030734
- Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social cognition*,

- 19(6), 625–666. doi: 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. 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
- 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), e0129601. doi: 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 (Washington, D.C.)*, 14(5), 950–961. doi: 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. doi: 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*. doi: 10.1080/08870446.2019.1606221
- van der Linden, W. J. (2007). A hierarchical framework for modelling speed and accuracy. *Psychometrika*, 72(3), 287–308. doi: 10.1007/s11336-006-1478-z
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204. doi: 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 depres-

- sion and anxiety disorders: The NESDA study. *PLoS ONE*, 11(11). doi: 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*, 1-10. doi: 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.
- 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). doi: 10.3389/fpsyg.2018.01902
- Zeidan, A., Khatri, U., Aysola, J., Shofer, F., Mamtani, M., Scott, K., ... Lopez, B. (2019). Implicit bias education and emergency medicine training: Step one? awareness. *AEM Education and Training*, 3(1), 81-85. doi: 10.1002/aet2.10124

Appendix

This appendix presents the R code used for obtaining the Rasch model and the log-normal model estimates from (Generalized) Linear Mixed-Effect models, respectively.

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`, `bf1`).
- `latency`: Column containing the latency of the IAT responses. Latency can be expressed in seconds or milliseconds. 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 (rows) for each subject must correspond to the total number of trials in the two associative conditions. For instance, in the IATs used for this study, respondents 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 is set at 0, so that the estimates of the fixed 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 2 (Table 1) of the estimates of the stimuli are centered at 0 (argument $(1|\text{stimuli})$), while in Model 3 (Table 1) respondents estimates are centered at 0 (argument $(1|\text{subject})$). In Model 1, the Null model, both stimuli and respondents are centered around 0.

The Rasch and log-normal estimates were obtained by means of the `lme4` package (Bates et al., 2015) in R. The `lme4` package can be installed and loaded with the following code:

```
install.packages("lme4") # install package
library(lme4) # upload the package for the estimation of
# the models
```

A Accuracy models specification

The code for the specification of the accuracy models is illustrated.

A.1 Model estimation

Model 1: The between-subjects variability is specified as random intercepts (i.e., $(1|\text{subject})$). The between-stimuli variability is specified as random intercepts (i.e., $(1|\text{stimuli})$) as well.

```
a1 <- glmer(correct ~ 0 + condition + (1|stimuli) + (1|subject),
  data = data, # IAT data in long format
  family = "binomial")
summary(a1) # summary of the results
```

Model 2: The between-subjects variability is specified as random intercepts centered around 0 (i.e., $(1|\text{subject})$). The within-stimuli between-conditions variability is specified as the random slopes of the stimuli in the conditions (i.e., $(0 + \text{condition}|\text{stimuli})$).

```
a2 <- glmer(correct ~ 0 + condition + (1|subject) +
  (0 + condition|stimuli), # stimuli random slopes
  data = data,
  family = "binomial")
```

```
summary(a2)
```

Model 3: The between–stimuli variability is specified as random intercepts, centered around 0 (i.e., $(1|stimuli)$). The within–subjects between-conditions variability is specified as the random slopes of the respondents in the conditions (i.e., $(0 + condition|subject)$).

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

A.1.1 Model comparison

Once the three models have been estimated, they can be compared with each other.

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

Since Model a2 and Model a3 have the same degrees of freedom, the χ^2 statistics obtained from their comparison is meaningless and cannot be used as a means for choosing the best fitting model. Comparative fit indexes should be used instead. The use of function `anova()` is just for the convenience of having all models comparative fit indexes, deviance, log-likelihood and degree of freedom on the same page.

A.2 Rasch model parameters

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

Model 1 results in overall respondents' parameters and overall stimuli parameters. Respondents overall ability parameters can be extracted and stored in a data frame:

```
ability <- data.frame(
  subject = rownames(coef(a1)$subject), # Respondents' ID
  ability = coef(a1)$subject[, 1] # Select the first column
)
```

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

```
easiness <- data.frame(
  stimuli = rownames(coef(a1)$stimuli), # Stimuli labels
  easiness = coef(a1)$stimuli[, 1] # Select the first column
)
```

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

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

Respondents overall ability parameters 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 condition-specific respondents parameters and overall stimuli parameters. Respondents' condition-specific ability parameters can be extracted as follows:

```
cond_ability <- coef(a3)$subject[, -1] # drop the first column
# (fixed intercept set at 0)
# rownames are the subjects' IDs
```

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

```
easiness <- data.frame(
  stimuli = rownames(coef(a3)$stimuli),
  easiness = coef(a3)$stimuli[, 1] # select only the
# random intercept estimates
)
```

B Log-time models specification

The code for the estimation of the log-normal models is the same as the one used for 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, only the code for the estimation of Model 3 and the code for extracting the log-normal model estimates is reported.

Model 3 can be estimated as follows:

```
t3 <- lmer(log(seconds) ~ 0 + condition + (1|stimuli) +
           (0 + condition|subject),
           data = data,
           REML = FALSE) # Maximum Likelihood estimation
summary(t3) # summary of the results
```

For the comparison of the log-time models, the same code as the one used for the comparison of the accuracy models can be used. The names of the models have to be changed accordingly, in this case from $\mathfrak{a}$ to $\mathfrak{t}$.

B.1 Log-normal model parameters

We report the code for extracting the log-normal model parameters for log-time Model 3, assuming it was the best fitting model according to model comparison. The same code used for extracting the parameters 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 can be obtained as follows:

```
cond_speed <- coef(t3)$subject[, -1] # drop the first column  
# (fixed intercept set at 0)  
# rownames are the subjects' IDs
```

Stimuli overall time intensity parameters can be obtained as follows:

```
intensity <- data.frame(
  stimuli = rownames(coef(t3)$stimuli),
  intensity = coef(t3)$stimuli[, 1] # select only the
) # random intercept estimates
```