University of Padova
Department of Philosophy, Sociology, Education, and Applied Psychology (FISPPA)

Ph.D. Course in Psychological Sciences (XXXIII Cycle)

Inglorious Measures: A Linear Mixed-Effects Model approach for modeling implicit measures data within a Rasch framework

Advisor: Prof. Egidio Robusto

Co-Advisor: Prof. Gianmarco Altoè

Ph.D. Candidate: Ottavia M. Epifania

Academic Year: 2019/2020

Muchos años después, frente al pelotón de fusilamiento, el coronel Aureliano Buendía había de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo.

A Francesco Epifania,
che ha trovato il modo di scampare alla
lettura di questa Tesi.

Contents

Preface **ix**

1 Psychological implicit assessment **1**

1.1 Automatic and controlled processes 1

1.2 The Implicit Association Test 5

1.2.1 Fields of application 8

1.3 The Single-Category Implicit Association Test 12

1.4 Fully-crossed design 15

1.4.1 More than one implicit measure 22

2 Typical scoring of implicit measures **25**

2.1 The IAT *D* score 26

2.2 The SC-IAT *D* score 28

2.3 A fairer comparison 28

2.3.1 Method 30

2.3.2 Data analysis 32

2.3.3 Results 35

2.3.4 Final remarks 42

2.4 R development 44

2.4.1 DscoreApp 44

2.4.2 `implicitMeasures` 51

3 Formal modeling **55**

3.1 Multinomial Models . . . 55

3.1.1 The Quad Model . . . 55

3.1.2 The ReAL Model . . . 62

3.2 Time and Accuracy models . . . 67

3.2.1 The Diffusion Model . . . 67

3.2.2 The Discrimination-Association Model . . . 71

3.3 Rasch Modeling . . . 73

3.4 Common features, advantages, and drawbacks . . . 77

4 Rasch model, Log-normal model, and LMMs **81**

4.1 Modeling dichotomous responses . . . 82

4.1.1 The Rasch model . . . 84

4.2 Modeling time responses . . . 91

4.2.1 The log-normal model . . . 92

4.3 Linear Mixed Effects Models . . . 94

4.3.1 Generalized Linear Mixed-Effects Model and Rasch Model . . . 95

4.4 Random structures . . . 99

4.4.1 GLMMs . . . 100

4.4.2 LMMs . . . 103

4.5 Other random structures . . . 107

5 Applications of (G)LMMs to IAT data **111**

5.1 Race IAT . . . 112

5.1.1 Method . . . 112

5.1.2 Data analysis . . . 113

5.1.3 Results . . . 116

5.1.4 Final remarks . . . 121

5.2 Chocolate IAT . . . 122

5.2.1 Method . . . 123

5.2.2 Data analysis . . . 124
5.2.3 Results . . . 126
5.2.4 Final remarks . . . 133

6 Multiple implicit measures: Models specification 137
6.1 Single measures models . . . 138
6.2 Comprehensive models . . . 139
6.2.1 Comprehensive GLMMs . . . 140
6.2.2 Comprehensive LMMs . . . 143

7 Multiple implicit measures: Empirical applications 147
7.1 Method . . . 147
7.1.1 Data analysis . . . 148
7.2 Results . . . 150
7.2.1 Single measures models . . . 151
7.2.2 Comprehensive models . . . 156
7.2.3 Relationship between model estimates and typical scoring . . . 158
7.2.4 Prediction of a behavioral outcome . . . 162
7.3 Final remarks . . . 164

8 In the end 167
8.1 The sound path . . . 167
8.2 The fair path . . . 179
8.3 In the end . . . 184

A Appendix A 187
A.1 Accuracy models specification . . . 189
A.1.1 Model estimation . . . 189
A.1.2 Rasch model parameters . . . 190
A.2 Log-time models specification . . . 191

A.2.1 Log-normal model parameters 192

B Appendix B **193**

B.1 Accuracy models specification 195

B.1.1 Model estimation 195

B.1.2 Model comparison 196

B.1.3 Rasch model parameters 196

B.2 Log-time models specification 197

B.2.1 Log-normal model parameters 198

References **199**

Preface

The advent of measures able to infer mental processes from the speed of respondents to computerized categorization tasks opened the access to processes that lie beyond people's awareness, but that can still influence their attitudes and social behaviors. These measure go under the name of implicit measures, and their use became more and more popular in social sciences, also thanks to the availability of accessible software for the administration of computerized categorization tasks. Despite the popularity implicit measures gained throughout the past decades, a lot of work still needs to be done to find a psychometrically sound approach to their modeling.

Usually, implicit measures are scored by averaging the response times across stimuli to obtain respondent-specific scores employed in further analyses. This approach has the clear advantage of being extremely easy and to provide a clear and interpretable measure of the implicit construct under investigation. However, the systematic variability between the stimuli, as well as the variability between the observations on the same respondent, are overlooked. These sources of uncontrolled error variance may generate statistically significant mean results that cannot be replicated when different samples of respondents and/or stimuli are used (Judd, Westfall, & Kenny, 2012). Given the replicability crisis that has been hitting psychology, and specifically social psychology, from the past few years, the need for more sound, accurate, and reliable analyses of data sets obtained with typical social psychology experiments (e.g., implicit measures) is of the uttermost importance.

The main objective of the Thesis is to provide new methods for more rigorous analyses of implicit measure data. In the long run, the repercussions of more rigorous data analyses

can be observed in the replicability of the results. For pursuing this aim, three paths are followed, one for a more sound approach to implicit measures data (sound path), one for a fairer comparison between implicit measures (fair path), and one for an easier (and more rigorous) way to compute implicit measure scores (easy path).

The sound path constitutes the main part of the Thesis. It is an attempt at finding new approaches for the analysis of implicit measures data. This is done by combining a classic of Psychometric Theories, the Rasch model, with a Linear Mixed-Effects Model approach. The focus is mostly on one of the most popular, used, and studied implicit measures, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), and on its single category version, the Single Category IAT (SC-IAT; Karpinski & Steinman, 2006). Their accuracy and time responses are modeled separately with distinct models. Consequently, the parameters either explain the processes leading to the accuracy responses or those leading to the time responses. The relationship linking these parameters can be explained and understood at a second (higher) level of modeling (van der Linden, 2009).

Traditionally, Item Response Theory and Rasch modeling treat items (stimuli) as fixed factors (i.e., unknown constants that do not vary as a function of the observational units), while respondents are treated as random factors (i.e., effects that vary according to the observational units, drawn from a larger distribution) (De Boeck et al., 2011). In this work, a slightly different approach was followed, also grounding on the data structure characterizing implicit measures. The fully-crossed design characterizing the IAT (see Chapter 1.4) allows one to conceptualize the stimuli as a manifestation of the super-ordered category they represent. Consequently, the specific set of stimuli used in an IAT is just one the possible set of stimuli that can be drawn from the same population of stimuli. Following this line of reasoning, it makes more sense to consider the stimuli as random factors and to treat them as random effects to make inferences on the larger population to which they belong, than to treat them as fixed factors.

Besides being a statistically more sound approach, acknowledging for the sampling variability of the stimuli implies that each stimulus has a potentially different functioning, and, consequently, a different impact on the observed responses. Therefore, if stimuli are treated

as random and their random variability is accounted for, it is possible to exploit it for the best to gather all the information they convey, to investigate their functioning, and their impact on the observed responses (Wolsiefer, Westfall, & Judd, 2017).

Linear Mixed-Effects Models allow for considering both respondents and stimuli as samples drawn from larger populations (and hence treating both of them as random factors) at the same time, resulting in more detailed and generalizable information at both levels.

Despite its wide use, the IAT is not the only available implicit measure and sometimes its use is not in line with one's aims. Given its structure, the IAT always results in a relative measure of the preference towards one target object contrasted to its (alleged) opposite. However, there are cases in which the object under investigation does not have a "natural" category to which it can be contrasted to. There might be also cases in which the focus is not on the relative preference but on the absolute positive or negative evaluation of one object.

In these occurrences, the IAT is not able to provide the measure of interest. The SC-IAT (Karpinski & Steinman, 2006) is often used as an alternative to the IAT when the aim is to obtain an absolute measure towards one object. The SC-IAT procedure results from a direct modification of the IAT one, where one of the target objects is dropped. Not infrequently, the IAT and the SC-IAT are administered together to obtain both a comparative and absolute evaluation of different attitude objects. By exploiting the flexibility of Linear Mixed-Effects Models, a comprehensive modeling of multiple implicit measures within a Rasch approach is possible and can be used for gaining more reliable and comparable estimates at both the respondents and stimuli levels, for each implicit measure.

However, the use of Linear Mixed-Effects Models for the conjoint analysis of multiple implicit measures within a Rasch framework is not a common approach. Effect size measures are the most popular and used scoring procedures for the IAT and the SC-IAT, referred to as *D* scores. These are often employed for comparing the performance of the IAT and of the SC-IAT on several variables used as criteria, such as the prediction of behavioral outcomes. The scoring procedures of both the IAT and the SC-IAT are affected by several artifacts, the most outstanding one being the lack of control on the sources of random variability in the data. Additionally, the IAT and the SC-IAT scoring and administration procedures present minor

differences, such as the inclusion of a response time window or not, that might still influence the comparison in their predictive ability. Taken together, the differences between the procedures potentially end up in misleading results. The fair path is an attempt at providing scoring methods for a fairer comparison between the IAT and the SC-IAT in terms of their capacity of predicting behavioral outcomes. New scoring algorithms for the IAT and the SC-IAT are introduced in the attempt of minimizing (non-necessary) procedural differences potentially affecting the comparison between the two measures. The procedures with which effect size measures are computed cannot overcome the issues of the sources of random variability characterizing implicit measures data. However, by aligning the differences in the procedures for scoring the IAT and the SC-IAT, the new alternatives should at least provide a means for a fairer comparison between the IAT and the SC-IAT. Consequently, the new, aligned, scoring algorithms produce (potentially) more reliable results regarding the comparison between the two measures on different criteria, such as the prediction of behavioral outcomes.

Finally, the easy path is oriented at providing open source and easy-to-use tools for the computation of the IAT and the SC-IAT scores. By automating the computational procedure and providing it open source, computational mistakes are prevented, the algorithms always end in the same results, which can be easily and openly replicated. In the long term, this would help for the replicability of the results obtained with implicit measures.

The structure of the thesis is outlined.

In Chapter 1, brief definitions of automatic and controlled processes are provided, and the main theoretical frameworks that have been proposed for conceptualizing the distinction between the two processes are outlined. The description of the IAT follows, along with the results of a literature review where the IAT use in different fields of application was investigated. The description of the SC-IAT is provided in Chapter 1 as well. The chapter ends with a description of the fully-crossed design characterizing implicit measures, and with the reasons why this structure might undermine the replicatibility of the results if it is not correctly accounted for.

Both the fair and easy paths are presented in Chapter 2. The typical and modified scoring procedures of the IAT and the SC-IAT are illustrated. Usually, the comparison between the

IAT and the SC-IAT is based on their predictive ability of behavioral outcomes, and the IAT tends to outperform the SC-IAT. The alignment of the administration procedure of the IAT and Sc-IAT, as well as of their scoring algorithms, should provide a comparison between the predictive ability of the two measures more centered on the implicit measures themselves than on the differences ascribable to the scoring and/or administration procedure.

The results of an empirical study where the predictive ability of the typical scoring procedures and that of the modified scoring procedures were compared are reported. Regardless of the algorithms used for scoring the implicit measures, the measure obtained from the IAT always outperformed the one obtained from the SC-IAT.

The easy component of Chapter 2 is composed of the presentation of two open source alternatives for the computation of the IAT and the SC-IAT typical scoring procedures. One of them is a Shiny app (i.e., DscoreApp; Epifania, Anselmi, & Robusto, 2019) for the computation of the IAT *D* score, while the other is an `R` package for the computation of the IAT and the SC-IAT *D* scores (the `implicitMeasures` package; Epifania, Anselmi, & Robusto, 2020c). DscoreApp was developed with the aim of providing researchers using the IAT an open source tools able to make the *D* score computation easier, without requiring for any programming experience. Moreover, DscoreApp also fosters the replicability of the results by providing a clear labeling and description for each scoring algorithm to which researchers can refer to. Additionally, the replicability of the results is undermined by the many steps that are required for cleaning and preparing the data (Ellithorpe, Ewoldsen, & Velez, 2015). By automating the procedure and providing clear labels and descriptions for the identification of each scoring algorithm, these errors should be prevented, and the results replicability should be enhanced.

DscoreApp presents two main shortcomings. One of them is an intrinsic limitation of Shiny apps. Since the code is put into the shiny interface, it is not possible to call it and run it from the command line, hence making it impossible to reproduce. While this might not constitute a problem for the average users, it is indeed a huge issue in an open science framework, according to which all the codes used for the analyses should be accessible at any time. Nevertheless, this issue can be overcome by storing the code in a public repository,

such as GitHub, as it was done for DscoreApp. Another important issue is that DscoreApp only computes the score for the IAT.

The `implicitMeasures` package is an `R` package developed for overcoming the two main limitations of DscoreApp. The package also comes with functions for cleaning the data sets of both the IAT and the SC-IAT and for plotting their results at either individual level or sample level.

Chapter 3 provides an overview of the main modeling frameworks that have been introduced for modeling IAT data. These frameworks can be distinguished according to the type of responses used for the estimation of the parameters. The Quad model (Conrey, Gawronski, Sherman, Hugenberg, & Groom, 2005) and the ReAL model (Meissner & Rothermund, 2013) are based on accuracy responses, while the Diffusion model (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007) and the Discrimination-Association model (Stefanutti, Robusto, Vianello, & Anselmi, 2013) account for both accuracy and time responses. Regardless of the type of responses they consider, these models are able to disentangle the most automatic processes from the most controlled ones intervening during the performance at the IAT. A common finding of these models is that the automatic associations are just one of the possible processes intervening during the performance at the IAT, and that other controlled processes, such as the recoding of the stimuli (ReAL, Diffusion Model) or the suppression of the automatically activated response (Quad model), play an important role as well. Despite their usefulness for the disentanglement of the IAT effect, these models come with some limitations. Most importantly, none of them can provide a detailed information at the level of the individual stimulus. This is a crucial point, also given that previous studies highlighted the importance of stimuli selection for a correct functioning of the IAT (e.g., Bluemke & Friese, 2006). Moreover, they overlook the fully-crossed structure of the IAT.

A Rasch modeling of the IAT does provide a detailed information on the stimuli functioning. By pinpointing the stimuli that give the highest contribution to the IAT effect, it is possible to delve deeper on the automatic associations driving the IAT effect, and hence to have a better understanding of the measure itself. However, the applications of the Rasch model to the IAT data performed so far are not save from criticisms. The most outstanding

one is related to the discretization of the time responses, which might cause a large loss of information. Moreover, also the Rasch modeling does not account for the random noise in the data due to the different sources of variability in the IAT data, which brings sources of dependency that are very likely breaking the local independence assumption.

An introduction to the Rasch model is provided in the first section of Chapter 4. The limitation of the Rasch model when it comes to its application to complex data structures, such as that of the IAT, and its similarities with the structure of Generalized Linear (Mixed-Effects) Models are presented as well. Given that the Rasch model is equivalent to a Generalized Linear Model (GLM) with a *logit* link function (i.e., the natural link function for binomial responses), the model matrix of the GLM can be extended to include the random effects able to address the sources of variability in the IAT data. This allows for obtaining Rasch model estimates from IAT data by employing Generalized Linear Mixed-Effects Models (GLMMs). The use of GLMMs for estimating Rasch model parameters accounts for the sources of random variability generating local dependence at the trial levels, hence resulting in more reliable estimates of the model parameters.

The log-normal model is introduced in the first section of Chapter 4 as well. By considering the normal density distribution of the log-time responses, the log-normal model allows for obtaining a parametrization of the data analogous to that provided by the Rasch model. Therefore, the discretization of the time responses needed for the application of the Many Facet Rasch Model (Chapter 3) can be avoided. The estimates of the log-normal model parameters can be obtained by applying Linear Mixed-Effects Models (LMMs) to the IAT log-transformed time responses.

The estimates of the Rasch model and the log-normal parameters do not directly result from the application of the (G)LMMs to either the accuracy responses or the log-time responses. They are obtained by adding the marginal modes of each level of the random factors (Best Linear Unbiased Predictors, BLUP) to the estimated fixed effects. The specification of models with different random structures allows for obtaining information at different levels of granularity on either the respondents or the stimuli.

The second section of Chapter 4 presents the specification of models with different ran-

dom structures for a meaningful Rasch and log-normal analysis of the IAT data. Three models for accuracy responses and three models for log-time responses are specified for obtaining the estimates of Rasch model and those of the log-normal, respectively. Besides the assumption on the distribution of the error term, the random structures of the accuracy and the log-time models are the same. The error term for the accuracy responses is modeled by assuming a logistic distribution, while the one for the log-time responses is supposed to follow a normal distribution. The random structures of the models are ordered according to their complexity, with the first one being the simplest one (i.e., Null model). The second and third models do have the same degree of complexity. They differ from each other according to the random factor on which they allow for the multidimensionality of the error variance, either the stimuli or the respondents.

Two empirical applications of the models presented in the second section of Chapter 4 are illustrated in Chapter 5. The first application was aimed at investigating the validity of the proposed models for the analysis of IAT data. To pursue this aim, a Race IAT was employed and the relationship between the estimates obtained from the Rasch and the log-normal models and the typical IAT scoring was investigated. By obtaining condition–specific stimuli estimates of the Rasch model, it was possible to investigate the contribution given by each stimulus to the IAT effect, resulting in a better understanding of the measure itself and in the identification of the malfunctioning stimuli that should be replaced or removed. The condition–specific respondents' estimates of the log-normal model, combined with the overall respondents' estimates of the Rasch model, brought further evidence in favor of the speed-accuracy trade-off and allowed for a better understanding of the IAT measure as expressed by the typical scoring algorithm.

The second application was aimed at understanding whether the estimates provided by the proposed modeling framework do result in a better inference of the implicit construct under investigation. As such, it is expected to lead to a better prediction of behavioral outcomes than the one given by the typical scoring procedure of the IAT. The second application was also aimed at testing the usefulness of the condition–specific stimuli estimates. If the stimuli estimates truly allow for pinpointing the most informative stimuli, as well as the least

informative ones, a higher amount of information should be obtained by selecting only the most informative stimuli. A smaller but highly informative data set can be obtained. The *D* score computed on the reduced data set should be more reliable than the one computed on the entire data set, and it potentially results in a better prediction of behavioral outcomes. An IAT for the implicit assessment of the preference for Dark or Milk chocolate (Chocolate IAT) was employed for pursuing these aims.

The Rasch model and the log-normal estimates did result in a better inference of the implicit preference, which in turn led to a better prediction of the behavioral outcome than the one provided by the typical scoring procedure. Moreover, the information on the contribution of each stimulus to the IAT effect allowed for pinpointing the most informative stimuli and for reducing the across-trial variability. The *D* scores computed on the reduced data set did result in a better prediction than those computed on the entire data set. Interestingly, even the *D* score computed on a reduced data set obtained by selecting only the least informative stimuli provided a better prediction than the one computed on the entire data set. These results pointed at the sensitivity of the *D* score to the across-trial variability. The reduction of the across-trial variability by selecting a smaller pool of stimuli leads to *D* score more related with external variables, even when the selected stimuli are the least informative ones.

The typical scoring methods of both the IAT and the SC-IAT have been presented, and their predictive ability in respect to a behavioral outcome has been investigated and compared with that provided by new scoring methods (Chapter 2). The new scoring methods do allow for a fairer comparison between the IAT and the SC-IAT, pointing at a better predictive ability of the IAT. However, the approach used in Chapter 2 has a main, outstanding fallacy, that is, the *post-hoc* separation of implicit measures administered concurrently to the same respondents.

When multiple measures are administered concurrently, each of them comes with its peculiar data structure and its method variance. Additionally, other sources of dependency have to be expected, namely the within–respondents between–measures variability. Moreover, since usually different implicit measures employ the same set of stimuli, also the within–stimuli between–measures variability should be expected. Therefore, on top of the method

specific variance of each measures, also other sources of variability should be taken into account to obtain reliable estimates.

Chapter 6 presents a comprehensive approach to the modeling of multiple implicit measures administered concurrently. The chapter firstly introduces the use of the models already presented in Chapter 4 for the separate modeling of the IAT and the SC-IAT. Despite this approach overlooks the within–respondents between–measures variability, it should still result in more reliable estimates than the *D* score. However, the estimates from the application of distinct models are not directly comparable between each other. Consequently, it is not possible to compare respondents' performance between implicit measures. The extension of the models to account for other sources of variability, hence allowing for the inclusion of multiple implicit measures in the same model, is illustrated.

An empirical application of the modeling approach in Chapter 6 is presented in Chapter 7. Data are the same as those in Chapter 2, hence including one IAT and two SC-IATs. The IAT and the SC-IAT data have been modeled separately with the (G)LMMS of Chapter 4 for obtaining the Rasch model and the log-normal estimates from each of them singularly. This was done for mainly two reasons. Firstly, to investigate the soundness of the proposed approach for modeling measures other than the IAT. Secondly, to investigate whether and how model estimates change if the within–respondents between–measures variability and the within–stimuli between–measures variability are not accounted for.

Results pointed out that, just by accounting for the method specific variance of each implicit measure, it is possible to obtain estimates that are more reliable than the *D* score, as it can be inferred from their better prediction of a behavioral outcome. Nonetheless, by analyzing the data from each implicit measure separately, the estimates at the stimuli level might be misleading (e.g., it is not possible to rule out whether the different functioning of the stimuli between measures is ascribable to an actual different functioning or to uncontrolled error variance). Moreover, the estimates at the respondents' level cannot be compared between implicit measures. The estimates obtained from the comprehensive modeling are similar to those obtained with the separate modeling of each measure. However, the comprehensive modeling allows for directly comparing the estimates at the levels of both respondents and

stimuli. Consequently, a better understating of the functioning of each implicit measure is obtained, and more meaningful inferences can be made. Besides a better prediction of the behavioral outcome than that provided by the typical *D* score, the estimates obtained from both the single modeling of implicit measures and those obtained from their comprehensive modeling allow for highlight the contribution of one of the SC-IATs to the prediction of the behavior. The contribution of the SC-IAT to the prediction of the behavior was completely lost when the typical scoring methods were used.

Finally, Chapter 8 summarizes the findings of all other chapters, and draws general conclusions based on the evidence reported in all the studies.

Chapter 1

Common measures for implicit psychological assessment

In this introductory chapter, a brief definition of automatic and controlled processes is provided, along with a summary of the main theoretical frameworks concerning the distinction of these processes.

The Implicit Association Test (IAT; Greenwald et al., 1998) is then presented, along with an overview of its use from the year of its first introduction (1998) to current days. The main fields of application of the IAT are illustrated as well. The issues related to the comparative measure that is gathered from the IAT is addressed by presenting an implicit measure able to provide an absolute evaluation towards one target object, namely the Single Category Implicit Association Test (SC-IAT; Karpinski & Steinman, 2006). The chapter ends with the illustration of the fully-crossed structure that characterizes the IAT and SC-IAT data.

1.1 Automatic and controlled processes

Throughout the past decades, social sciences have seen a growing interest in the possibility of assessing people's attitudes, preferences, opinions, personality traits and other psychological constructs without directly asking them but by inferring them from respondents' performance to computerized categorization tasks (i.e., implicit measures). Usually, implicit measures are

tasks in which respondents' are called to sort stimuli representing different categories. The stimuli are specifically selected to trigger the activation of implicit processes, which are defined as processes operating outside of people's awareness but that can still affect behaviors, decisions, and social judgments (Greenwald & Banaji, 1995; Greenwald & Lai, 2020). The use of response times for inferring mental processes activated by a stimulus has a long tradition in psychology (see Greenwald & Lai, 2020). The implicit investigation of psychological and social constructs has been now widely recognized and it earned the label of "implicit social cognition" (Greenwald & Banaji, 1995).

According to dual process theories (e.g. Devine, 1989; Fazio & Olson, 2003), two distinct but mutually reliant processes are involved in people's social behaviors and attitudes. This implies that they are different manifestations deriving from the same single–representation. Automatic and controlled processes usually happen simultaneously and involve automatic and controlled components. Differently from implicit processes, which do not require the availability of cognitive resources for being activated, controlled processes do require a cognitive effort, resulting from the interaction between the person's willingness to engage in that process and the availability of cognitive resources and time for engaging in the process (Fazio & Olson, 2003). Moreover, controlled processes allow one to compare new information with previous experience and to make inferences on the environment in which they occur, making them more sensitive to social judgment and social desirability (e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

Controlled processes are usually assessed by means of what are defined direct measures, such as self-report scales. Consistently with the definition of controlled processes reported so far, direct measures assess the construct under investigation by directly giving an instruction to report it, presuming a certain degree of awareness of the construct itself (Greenwald & Lai, 2020). Conversely, automatic processes are assessed with measures that assume no introspective awareness of the construct itself, and, most importantly, they measure the construct of interest without directly asking to report it (Greenwald & Lai, 2020).

A common trend for concurrently assessing controlled and automatic processes underlying psychological constructs is to administer a direct measure for capturing the former ones

and an indirect measure for capturing the latter ones (Brownstein, Madva, & Gawronski, 2019).

The controlled components of psychological constructs assessed by direct measures have been found to be highly correlated with the automatic components of the same psychological constructs assessed by indirect measures (e.g., Nosek, 2007). However, this correlation is reduced when the psychological constructs under investigation involve socially sensitive topics, such as racial prejudice (Greenwald et al., 2009; Nosek, 2007). Another evidence pointing towards a dissociation between controlled and implicit processes comes from the type of behaviors predicted by these processes. Controlled processes assessed by direct measures have been found to be good predictors of deliberate behaviors (i.e., behaviors on which people have a certain degree of control), such as brand choice or food preference. Conversely, non-deliberate behaviors (i.e., spontaneous behaviors on which people do not have a control), such as prosemic and social distance in the interaction with members of stigmatized groups (e.g., Dovidio, Kawakami, & Gaertner, 2002), are better predicted by implicit processes assessed with indirect measures (e.g., Meissner, Grigutsch, Koranyi, Müller, & Rothermund, 2019; Perugini, 2005; Wilson, Lindsey, & Schooler, 2000). Different hypotheses have been formulated for explaining the worse predictive ability of implicit measure in respect to behavioral deliberate choices, such as food/brand choices. For instance, Greenwald and Banaji (2017) claim that the automatic associations assessed by indirect measures can trigger deliberate thoughts about the associations themselves but they are unlike to predict behaviors. Conversely, decisions and choices might be more ascribable to the deliberate thoughts triggered by the automatic associations. Meissner et al. (2019) ascribe the scarce predictive ability of implicit measures in respect to deliberate choices to the type of implicit processes that these measures are tapping. Indeed, by the way in which they are designed, indirect measures tap the *liking* component of the target objects, that is, how much an object is positively or negatively evaluated. Behaviors and choices are rather guided by a *wanting* component, that is, how much a target object is desired.

The apparent dissociation between automatic and controlled processes is still not solved, despite different theoretical frameworks have been trying to provide a conceptualization of

these processes that should be able to explain these peculiar patterns of associations with external criteria (Perugini, 2005).

According to dual process theories, it should not be surprising to observe such different patterns between controlled and automatic processes,and their predictive ability might change according to each and every kind of behaviors. Despite the two manifestations of the same attitude should provide a distinctive and unique prediction of the behavior, there might be cases in which one overrides the other in predicting the behavioral outcome. Moreover, controlled and automatic processes can be affected differently by other variables which are not directly accounted for by either the direct or indirect assessment, such as social desirability (direct measures) or a moment of distraction (indirect measures). Besides the external variables that can differently affect the direct and indirect assessment, the low correlation between the two measures of the same psychological construct can be due to the discriminant validity between two different types of measures, one based on introspection and on explicit and controlled responses, the other one based on response times and automatic responses.

Alternative explanations posit the coexistence of both explicit and implicit attitudes towards the same attitude objects (Dual Attitudes Model; Wilson et al., 2000). The combination between the degree of awareness of an implicit attitude and the extent to which motivation and cognitive resources are needed for overcoming the implicit attitude in favor of the explicit one generates different types of dual attitudes, namely repression, independent systems, motivated overriding, and automatic overriding. The first dual attitude, *repression*, requires the capacity of the explicit attitude to override the implicit one, which is kept outside of awareness. The dual attitude *Independent systems* requires the presence of two attitudes towards the same object, one within awareness (explicit), and the other one outside of awareness (implicit). However, in this case, both systems develop evaluations, and they influence different type of responses (i.e., explicit and implicit responses). In both dual attitudes *repression* and *independent systems*, the implicit attitude did not reach awareness. In *motivated overriding*, people are completely aware of the existence of the implicit attitude, but they are motivated to override it because they view it as illegitimate or they are ashamed of it. Clearly, this process requires the availability of cognitive resources for the explicit attitude to override the implicit

one. The difference between the *motivated overriding* and the *automatic overriding* lies in the automaticity of the overriding process. In the former case, it is a motivated and aware process, while in the latter case, the overriding occurs outside of people's awareness.

Two main assumptions underlie all dual attitudes. Firstly, once an implicit attitude is formed from previous experience, it does not require for cognitive resources to be activated, and it is activated every time the attitude object is encountered. Second, the explicit attitudes do require cognitive capacity and motivation to be retrieved. Both assumptions are in line with the conceptualizations of implicit and explicit processes presented so far. The independence assumed between implicit and explicit attitudes towards the same attitudes object allows for speculating a double dissociation pattern in the prediction of behaviors. It follows that implicit attitudes can solely predict behaviors that are not under people's awareness and control, while explicit attitudes are directly related to behavioral responses under people's awareness and control.

1.2 The Implicit Association Test

Several implicit measures have been introduced for tapping the implicit component of attitudes, preferences, self-esteem, and other psychological constructs, such as the IAT (Greenwald et al., 1998), the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), and the Sorting Paired Features task (SPF; Bar-Anan, Nosek, & Vianello, 2009), just to name a few. Nonetheless, the IAT is the implicit measure that presents the best psychometric properties when compared to other commonly employed implicit measures (Bar-Anan & Nosek, 2014). Moreover, by appropriately changing the labels of the attitude objects and leaving its structure unaltered, the IAT is easily adaptable for the investigation of a broad range of topics (Zogmaister & Castelli, 2006), such as stereotypes, attitudes, and self-concept. This characteristic of the IAT fostered its use in many different fields of applications, including law, criminal justice, education, marketing, and business (e.g. Epifania, Robusto, & Anselmi, 2020b; Greenwald et al., 2009; Greenwald & Lai, 2020).

The IAT measures the strength of the associations between concepts by considering the

speed and accuracy with which prototypical exemplars of two objects categories (e.g., *Coke* and *Pepsi* images in a Coke-Pepsi IAT) and of two evaluative dimensions (i.e., *Good* and *Bad* attributes) are sorted in the category to which they belong by means of two response keys.

The usual structure of an IAT (illustrated in Table 1.1) is composed of 7 blocks.

Table 1.1: Coke-Pepsi IAT structure (adapted from Greenwald et al., 2003).

Block	Function	Left Response key	Right Response key
B1	Pure practice	Good words	Bad words
B2	Pure practice	Coke	Pepsi
B3	Associative practice	Good & Coke	Bad & Pepsi
B4	Associative test	Good & Coke	Bad & Pepsi
B5	Pure practice	Pepsi	Coke
B6	Associative practice	Good & Pepsi	Bad & Coke
B7	Associative test	Good & Pepsi	Bad & Coke

Note: The order of presentation of Blocks B3 and B4 and blocks B6 and B7 are counterbalanced across respondents.

The labels of the four categories, both the evaluative ones and the target objects ones, are fixed at the top left and right corners of the computer screen. The stimuli are presented sequentially at the center of the screen.

The first two blocks are pure practice blocks in which respondents have to sort the stimuli belonging to either the evaluative dimensions (Block B1) or the stimuli belonging to the object categories (Block B2). These blocks have the purpose of letting the respondents familiarize with the stimuli and the task. Blocks B3 and B4 form the first associative condition. In these blocks, the object category *Coke* shares the response key with *Good* attributes, while the object category *Pepsi* shares the response key with *Bad* attributes (Coke/Good-Pepsi/Bad condition, CGPB). In Block B5, the labels of the object categories switch their positions on the computer screen. This block is a practice block that lets respondents familiarize with the

new locations of the labels. Blocks B6 and B7 constitute the contrasting associative condition, in which the categorization task is reversed. In these blocks, *Pepsi* and *Good* exemplars are sorted with the same response key, while *Coke* and *Bad* exemplars are sorted with the other response key (Pepsi/Good-Coke/Bad condition, PGCB). The assumption underlying the IAT functioning is that it is easier to sort together the exemplars of two categories when these categories are strongly associated with each other than when they are not. Consequently, respondents are supposed to perform better, in terms of faster time responses and higher accuracy, in the condition consistent with their automatically activated association(s).

The two associative conditions of the Coke-Pepsi IAT in Table 1.1 are depicted in Figure 1.1.

(a) Coke-Good/Pepsi-Bad condition (b) Pepsi-Good/Coke-Bad condition

Figure 1.1: Associative conditions of a Coke-Pepsi IAT.

During the administration of the IAT, respondents might be given feedback of their performance. If the IAT administration includes feedback, a red "X" appears every time a stimulus is sorted in the incorrect category. To proceed with the experiment, respondents have to correct their response. When the IAT does not include feedback in the administration, respondents are not notified when they commit errors, and they keep going with the experiment.

The so-called IAT effect results from the difference in respondents' performance between the two conditions, and it is usually interpreted by means of the *D* score (Greenwald et al., 2003, see Chapter 2.1).

1.2.1 Fields of application

A recent literature review (Epifania, Robusto, & Anselmi, 2020b) showed an increasing use of the IAT in wider and more varied fields of application. Since the year of its first introduction (1998), the IAT has been used in more than 1,400 studies, investigating different topics. By reading the abstracts of 1,418 papers citing and using the IAT (i.e., number of citations from 1998 to October 25th 2019, date of the search on Scopus database), it was possible to identify 6 main fields of application of the IAT: Social psychology ($n = 513$), Clinical and dynamic psychology ($n = 290$), Addiction ($n = 113$) , Food research ($n = 43$), Marketing research (n =34), and Other applications ($n = 425$). Figure 1.2 depicts the trend lines for each field of application from 1998 to 2019.

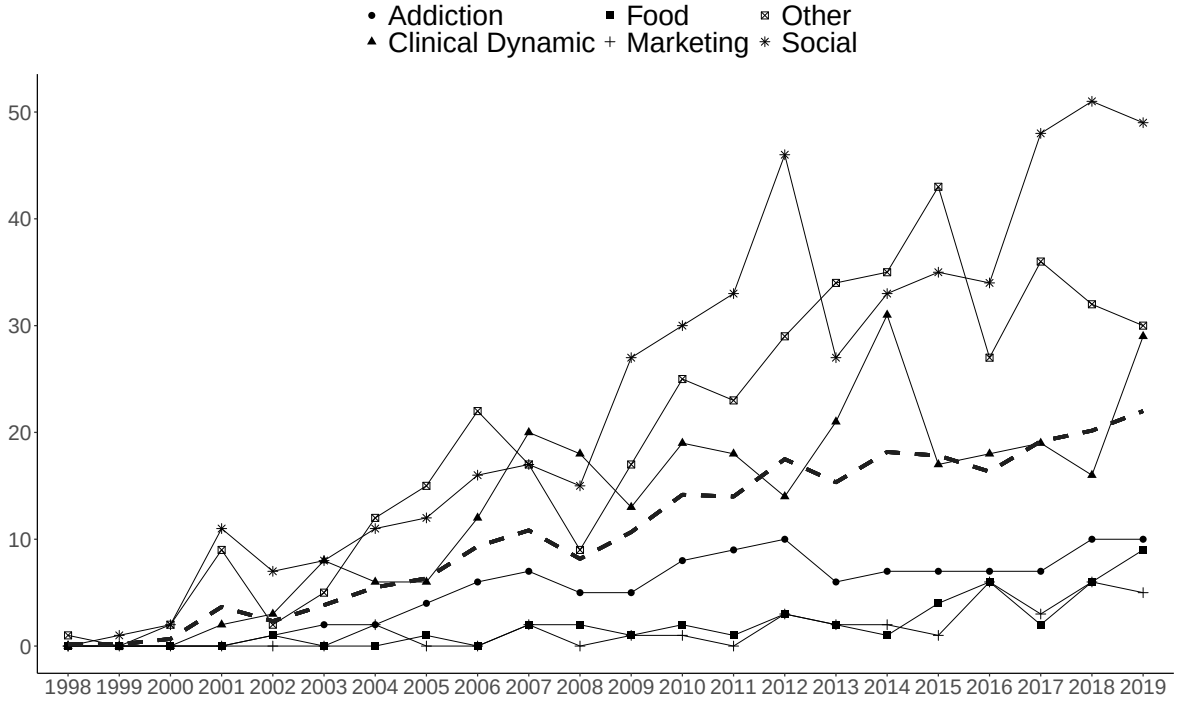


Figure 1.2: Trend lines of each field of application of the IAT throughout from 1998 to 2019.

The dashed line in Figure 1.2 represents the average trend of IAT use across the fields of applications. It clearly points at a constant and on-going growth in IAT use throughout the years.

Trend lines of Social psychology and of Other appear always above the mean trend line.

Clinical and dynamic psychology trend line is the most inconsistent one throughout the years. The trend lines of Food and Marketing are similar between each other, both pointing at a higher use of the IAT during the past few years.

Besides the general fields of applications of the IAT, it is interesting to delve deeper on the specific topics for which the IAT was employed (depicted in Figure 1.3).

Figure 1.3: Word cloud of the fields of application of the IAT. The bigger the word font, the more common the use of the IAT for the investigation of that specific topic.

Not surprisingly, the IAT was mostly used for investigating implicit stereotypes and attitudes, such as implicit racial prejudice, gender stereotypes, attitudes towards obese people and other social groups (i.e., psychiatric patients and people with disabilities).

The IAT has also been used for investigating addiction, unhealthy behaviors (i.e., smoking, alcohol consumption), personality traits, and self-perception.

The following paragraphs outline a brief summary of IAT use in each field of application. The complete report of the fields of application of the IAT, along with the samples used in the studies, can be found in Epifania, Robusto, and Anselmi (2020b).

IAT in Social psychology. The resistance of the IAT to self-presentation strategies (e.g., Greenwald et al., 2009) made it particularly appealing for investigating socially sensitive

topics, such as racial prejudice and, more generally, stereotypes and attitudes towards different out-groups. Studies on this topic are followed by studies focused on gender stereotypes, and on the investigation of illness-related attitudes. In this thesis, illness-related attitudes is a label used for indicating attitudes towards people with either mental or physical disabilities, psychiatric patients, people with HIV, cancer patients and day-care patients in general, and suicide survivals.

Despite with lower frequency, the IAT is also used to assess attitudes towards people professing different religions, towards non-native English speakers, and bias towards people with low incomes. Papers composed by multiple studies in which attitudes towards multiple out-groups (e.g., out-group prejudice and weight bias) were concurrently investigated were common. The IAT was also used to investigate topics related to the Stereotype Content Model (SCM; Fiske, Cuddy, Glick, & Xu, 2002), including infra-humanization, and to investigate the effectiveness of experimental manipulation to change/induce attitudes, even towards non-real groups.

IAT in Clinical and dynamic psychology. The IAT was mostly used for the implicit assessment of self-esteem. Personality traits (e.g., Big Five personality traits), anxiety, and personality and mood disorders according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnosis definition were fairly investigated as well. The IAT was also commonly used for the implicit assessment of suicidal tendencies, aggressiveness, emotions, and clinical sex behavior (e.g., pedophilia).

IAT in Addiction research. The Vast majority of studies using the IAT in this field were focused on the investigation of alcohol addiction, followed by studies on nicotine and smoking addiction. The concurrent investigation of multiple addictions, such as drinking with smoking or drinking with gambling, was quite uncommon.

IAT in Food research. The IAT was mainly used to investigate the preference for different kinds of food, the preference for healthy over unhealthy food, or food perception in general.

IAT was also employed for investigating attitudes towards dieting. Less common topics were food craving, food self–control, and the effect of the time of day on food preference.

The IAT in marketing research. Marketing research is one of the most recent fields of applications of the IAT. Most of the studies are focused on the implicit evaluation of the preference between different brands. Additionally, the IAT has been employed for studying the processes driving the decision to purchase some products, and the role of products labels and packaging in influencing the purchase of that specific product.

The IAT in macro-area Other. The studies included in this field of application cover a broad and extensive range of topics, from gender perception of odds and even numbers (Wilkie & Bodenhausen, 2015) to work related stress (Klein et al., 2012) and romantic attachment (Zayas & Shoda, 2005). Studies aimed at the validation of the IAT are included as well, and they compose the vast majority of studies in this field of applications. They are followed by studies on human perception and studies on methodology. The distinction between measure validation papers and methodology papers is quite subtle. Measure validation studies include papers aimed at the validation of the IAT procedure (e.g., Greenwald et al., 1998), its score (e.g., Greenwald et al., 2003), and the factors that may affect the IAT effect (e.g., Bluemke & Friese, 2006). Methodology papers includes studies in which existing formal models were used for modeling IAT data, such as the application of the Many-Facet Rasch Measurement Model (Linacre, 1989) in Anselmi, Vianello, and Robusto (2011) or the application of the Diffusion Model (Ratcliff, 1978) in Klauer et al. (2007). Studies aimed at the validation of *ad-hoc* models for IAT data, such as the Quad Model (Conrey et al., 2005) or the Discrimination-Association Model (Stefanutti et al., 2013) are included under the label methodology as well.

1.3 The Single-Category Implicit Association Test

The IAT has vastly proven its effectiveness and usefulness in providing a relative measure of the preference towards one object category compared with a contrasted one. However, the measure obtained from the IAT presents two main shortcomings.

Firstly, the relative measure provided by the IAT is not able to understand whether the performance is driven by a positive evaluation towards one of the objects, a negative evaluation towards the opposite one, or a combination of the two evaluations. Sticking with the Coke-Pepsi IAT example in Section 1.2, faster responses in the Coke/Good-Pepsi/Bad associative condition might be due to either a preference for Coke (faster responses in sorting *Coke* images and *Good* attributes with the same response key), a dislike for Pepsi (Coke) (faster responses in sorting *Pepsi* images and *Bad* attributes with the same response key), or even a combination of the two. One could try to decompose the IAT effect by computing separate scores for the trials in which *Good* attributes and *Coke* (*Pepsi*) images are associated and in which *Bad* attributes and *Coke* (*Pepsi*) images are associated. Even by doing so, it is not possible to obtain an absolute measure of the preference towards one of the two beverages (Nosek, Greenwald, & Banaji, 2005). The IAT is based on a comparative task and, as such, it can only result in a comparative measure. Consequently, the IAT is not the most appropriate choice when the focus in on the assessment of the absolute positive or negative evaluation of a single object. An appropriate example of a situation in which an absolute measure towards only one target object is preferable is self-esteem. In this case, the interest would be on how much a person values himself/herself. However, studies that employed the IAT for implicitly investigating this construct contrasted the category *Self* or *Me* with a generic category, like *Other* (e.g., Hiller, Steffens, Ritter, & Stangier, 2017) or *Not Me* (e.g., Fatfouta & Schröder-Abé, 2018). Consequently, the measure resulted in an indirect evaluation of how much a person valued himself/herself in comparison to others, and not in how much a person valued himself/herself (Karpinski & Steinman, 2006).

Secondly, since the IAT effect depends on the relative evaluation between the two categories, the choice of the contrasted category is of the uttermost importance. In some cases,

a clear contrasted category is not available, and the researcher has to make arbitrary choices. Sticking with the Coke-Pepsi IAT example in Section 1.2, the relative attitude towards Coke strongly depends on the attitude towards Pepsi (and vice-versa). Take for example a respondent who does not like Pepsi at all and that does not hold either a negative or positive evaluation towards Coke. By using Pepsi as the contrasted category of Coke, a strong positive IAT effect can be found, although it would be mostly due to a dislike for Pepsi and not to a true positive attitudes towards Coke. By replacing Pepsi with another soft drink, the resulting IAT effect might change, and even result in a negative score.

Different alternatives have been introduced to overcome the issue of the relativeness of the IAT measure, like the SC-IAT (Karpinski & Steinman, 2006).

The SC-IAT results from a slight modification of the IAT procedure. It is aimed at assessing the strength of associations between concepts by considering the speed and accuracy with which different stimuli are sorted in their reference categories. The assumption that underlies the functioning of the SC-IAT is the same as that underlying the functioning of the IAT, namely, that it is easier to sort together exemplars of two categories when they are strongly associated with each other than when they are not. However, differently from the IAT, the exemplars of only one object category (e.g., *Coke* in a Coke SC-IAT), along with the exemplars of two evaluative dimensions, are presented. The usual structure of a SC-IAT is illustrated in Table 1.2.

Table 1.2: Coke SC-IAT structure (adapted from Karpinski & Steinman, 2006).

Block	Function	Left Response key	Right Response key
B1	Associative practice	Good & Coke	Bad
B2	Associative Test	Good & Coke	Bad
B3	Associative practice	Good	Bad & Coke
B4	Associative Test	Good	Bad & Coke

Note: The order of presentation of Blocks B1 and B2 and Blocks B3 and B4 are counterbalanced across respondents.

The SC-IAT is usually composed of 4 blocks. In the first two Blocks (B1 and B2), the target object *Coke* and the *Good* attributes share the same response key, while *Bad* attributes are sorted with the opposite response key (Coke/Good-Bad condition, CG). In the last two blocks (B3 and B4), target object *Coke* is sorted with the same response key as *Bad*, while *Good* words are sorted with the opposite response key (Coke/Bad-Good condition, CB).

Blocks B1 and B3 are associative practice blocks, while Blocks B2 and B4 are the actual critical blocks that constitute the two associative conditions.

The two conditions of the Coke SC-IAT described in Table 1.2 are depicted in Figure 1.4.

(a) Coke-Good/Bad condition (b) Coke-Bad/Good condition

Figure 1.4: Associative conditions of a Coke SC-IAT.

The SC-IAT administration usually includes a response time window (rtw) at 1,500ms, after which the stimulus disappears and a warning message (e.g., "Respond more quickly!") is given to the respondent. Every correct response is signaled by a green "O", while every incorrect response is signaled by a red "X". Differently from the IAT, respondents do not have to correct their incorrect responses to go on with the experiment.

The rtw and feedback for every response are the two characteristics that differentiate the SC-IAT from the Single Target Implicit Association Test (ST-IAT; Wigboldus, Holland, & van Knippenberg, 2004). Nonetheless, the names (and procedures) of the two measures are used interchangeably (e.g., Bar-Anan & Nosek, 2014). The SC-IAT effect results from the difference in respondents' performance between the two contrasting conditions, and it is usually expressed by a modification of the IAT *D* score (Karpinski & Steinman, 2006, see Chapter 2.2).

1.4 Implicit measures fully-crossed design

Suppose that two respondents, Lara and Francesco, are presented with the Coke-Pepsi IAT in Section 1.2.

Lara might be, on average, more accurate (or faster) than Francesco. The difference in the overall performances of Lara and Francesco can be ascribable to their individual differences and characteristics. The between–respondents variability is the expression of the differences due to individual characteristics, across the associative conditions.

The set of exemplars chosen for representing each category presents their own variability as well. The *Coke* logo can be immediately recognized and sorted in its own category, while an image of an old fashioned can of Coke might be less familiar and hence might need more time for being recognized and sorted. Similarly, the attribute *evil* might be immediately recognized as belonging to the evaluative dimension*Bad*, while attribute *wicked* might not be immediately recognized as belonging to the evaluative dimension*Bad* [1]. The between–stimuli variability is the expression of the sampling variability due to stimuli characteristics, across associative conditions.

So far, only the between–respondents variability and the between–stimuli variability have been considered as the expression of individual differences and stimuli differences, respectively. The effect of the IAT associative condition has not been mentioned yet. The change in respondent' performance according to the IAT associative condition is the main object of investigation when an IAT is administered.

The respondents' individual differences might be exacerbated or diminished by the effect of the associative condition. Recall that Lara showed a better overall performance than Francesco. However, the difference in their performances might be attenuated in the Coke-Good/Pepsi-Bad condition. This attenuation might be due to several reasons. For instance, Lara might keep her performance unaltered because she is neither fond of Coke nor disgusted by Pepsi, while Francesco might show a better performance because he particularly likes

[1] The attribute *evil* is the English translation for *cattivo* (Italian). The attribute *wicked* is the English translation for *malvagio*. The spread Index (which varies from 0 to 1, where 1 indicates a high spread) for the former one is 0.85, while for the latter is 0.36 (Bambini & Trevisan, 2012)

Coke over Pepsi. In the opposite condition, the difference between the performances of Lara and Francesco might be exacerbated. Lara might still keep her performance unaltered, while Francesco might be struggle in associating his favorite soda with negative attributes. The within–respondents between–conditions variability is hence the expression of the variability in respondents performance that is ascribable to the effect of the associative condition.

Usually, to investigate whether the associative condition had an effect on respondents' performance, a *by-participant* approach is undertaken (an individual respondents score is obtained by taking the difference in the average response time computed across trials in each condition; Judd et al., 2012).

There are no reasons to suppose that stimuli are immune to the effect of the associative condition. Some of the stimuli might be more easily sortable in one associative condition than the other for a number of reasons, including the specific attitude of the respondents. The within–stimuli between–conditions variability indicates whether the stimuli functioning changes according to the associative condition in which they are presented. By exploiting this information, it is possible to obtain a measure of the contribution given by each stimulus to the IAT effect. Following this line of reasoning, the more (less) a stimulus functioning changes between condition, the higher (lower) its contribution to the IAT effect.

To investigate the effect of an experimental variable on the stimuli functioning, a *by-stimulus* approach is undertaken (an individual stimulus score is obtained by taking the difference in the average response time across respondents in each condition; Judd et al., 2012).

There is still one source of variability missing, that is, the variability due to the respondents' reactions to each stimulus. Lara is a better respondent than Francesco, but she also does not have a particular preference for any of the sodas. It can be speculated that Lara would generally react in the same way to each of the stimuli representing the two brand of sodas. On the other hand, Francesco has a strong preference for Coke. As such, he might be better than Lara in recognizing even the above-mentioned old-fashioned can of Coke. Consequently, he would have a better performance than Lara on this specific stimulus. The interaction between respondents and stimuli variability is hence the expression of the interaction between respondents and stimuli characteristics.

This practical example was meant to give an overview of how the sources of variability in the IAT data are generated. A more detailed illustration of the IAT structure, and its related sources of random variability, is further illustrated.

In the IAT[2], the respondents are presented multiple times with the stimuli nested in two levels of two independent variables, namely the evaluative dimensions (*Good* vs *Bad*) and the target objects (e.g., *Coke* vs *Pepsi*). In a multilevel modeling perspective, respondents and stimuli can be considered at the same level.

Another independent variable, at a higher level, is the associative condition, which includes both the respondents and the stimuli, and it is in turn composed of two levels (e.g., CGPB condition and PGCB condition). Usually, the interest is on how the respondents' performance vary according to the associative conditions, which are the independent variables focus of attention when analyzing IAT data.

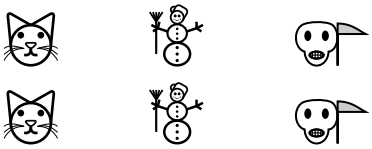
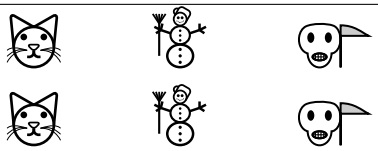
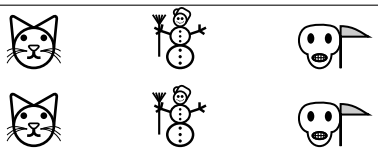
The stimuli representing each of the categories are presented multiple times both within and between the associative conditions. As such, besides being crossed with the respondents, the stimuli are crossed with the two levels of the associative condition. Similarly, the respondents are asked to respond to the stimuli multiple times both within and between the associative conditions. Therefore, besides being crossed with the stimuli, also the respondents are crossed with the associative conditions.

Summarizing, the respondents and the stimuli are crossed with each other, and they are both crossed with the associative conditions. This data structure is usually referred to as a fully-crossed design, since all levels of the independent variables are crossed with each other (Westfall, Kenny, & Judd, 2014). The responses x_{ps} of three participants p to three stimuli s (i.e.,  ,  , ) in the two IAT associative conditions c are represented in Table 1.3 to exemplify the fully-crossed structure characterizing the measure.

Each cell of Table 1.3 contains the unique combination respondent $\times$ stimulus for every repetition of the stimulus in each associative condition (x_{psc}). In other words, each cell contains the response to each IAT trial, that is the lowest level of observation, crossed with the stimuli, the respondents, and the associative conditions. Each stimulus is presented multiple

[2]The SC-IAT presents the same data structure, although one of the target categories is dropped.

Table 1.3: Fully-crossed design.

	Condition A			Condition B		
p_1						
p_2						
p_3						

Note: p: Respondents.

times in each condition to each respondent, hence multiple responses are observed for each of them. The illustration in Table 1.3 clearly represents how the dependency at the level of the single observation is generated by the random noise in the data.

Data deriving from a fully-crossed design, such as that of the IAT, should be carefully analyzed to account for the sources of variability, related to the participants, the stimuli, the associative conditions, and to their interaction (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013; Judd, Westfall, & Kenny, 2017; Westfall et al., 2014; Wolsiefer et al., 2017). These sources of variability generate dependencies between the observations that violate the assumption of conditional independence. This assumption is the basic assumption underlying data analysis in Social sciences, according to which, once the effect of the variability due to one latent variable is accounted for, the remaining variability can be explained by means of the experimental factors.

Usually, the IAT (or SC-IAT) effect is expressed by means of an effect size measure that is obtained by aggregating the responses across the trials in each associative condition, and dividing these quantities by the standard deviation computed on the pooled trials of both blocks. This measure is the so-called *D* score (Chapter 2; Greenwald et al., 2003; Karpinski

& Steinman, 2006), and it provides an easy-to-compute and easy-to-interpret measure of the implicit bias assessed by the implicit measure.

However, the easiness with which D score can be computed and interpreted comes with drawbacks that cannot be ignored. The computation procedure of the D scores implicitly entails that the stimuli are taken as being the exhaustive representation of the population of stimuli (i.e., fixed factors), while respondents are considered as just one of the possible samples that could be drawn from a population (i.e., random factors) (Judd et al., 2012). Considering the respondents as a random factor (hence treating them as random to make inferences on their reference population), and stimuli as a fixed factor, defines a *by-participant* analysis.

Considering stimuli as fixed factors has important consequences both theoretically and statistically. Firstly, treating stimuli as fixed factors assumes that they all have the same effect on the observed measure, or, in other words, they all have the same functioning. Moreover, the only inferences that are allowed concern the population from which the sample of respondents is drawn. The results are hence generalizable at the respondents' level but not at the stimuli level. Therefore, the replicability of the results to samples drawn from different populations is bounded to the use of the same exact set of stimuli (Judd et al., 2012). Finally, averaging across trials leaves the variability due the between–stimuli variation uncontrolled.

When the unaccounted error variance (the by-stimulus variation) is confounded with the effect of interest (the effect of IAT associative conditions), the risk of committing Type I error is inflated: The significant difference between the means can be due to the sources of error variance and not to the experimental effect (Barr et al., 2013; Judd et al., 2012; McCullagh & Nelder, 1989).

Not accounting for the sources of error variance also has important effect when the source is orthogonal (i.e., independent) to the effect of interest. In this case, the un-controlled error variance will reduce the power for testing the relevance of the effect of interest. Consequently, the importance of the experimental manipulation is underestimated (Barr et al., 2013).

In the IAT (SC-IAT) case, the investigation at the stimuli level is not aimed at the functioning of each individual stimulus. Rather, the aim is to gather the information provided by

each of the stimuli categories. The individual stimuli are assumed to be the realizations of the category to which they belong. As such, it makes sense to conceptualize the stimuli used in an IAT (SC-IAT) as possible samples drawn from the populations defined by the category to which they belong (Wolsiefer et al., 2017). Given this conceptualization, considering stimuli as fixed factors, and not accounting for their sampling variation, appears to be a great fallacy when treating IAT (SC-IAT) data. Moreover, when the by-stimulus variation is not accounted for, all the information that can be gathered from them is overlooked (Wolsiefer et al., 2017).

The between–stimuli variability can be accounted for by performing *by-stimulus* analyses (Judd et al., 2012). Differently from the *by-participant* approach (i.e., the respondents are treated as random factors and the stimuli are treated as fixed factors by averaging per participant across stimuli) presented so far, the *by-stimulus* approach treats the stimuli as random factors and the participants as fixed factors. The stimuli are hence assumed to be just one of the possible sets of stimuli that can be drawn from a population of stimuli. The overall scores of each stimulus can be obtained by averaging per stimulus across participants. Clearly, the same pitfalls highlighted for the *by-participant* approach apply for this instance, but reversed.

Not considering the between–participants variability affects the computation of the mean score for each stimulus, and the statistical tests that can potentially be performed. However, the *by-stimulus* analyses allow for the generalization of the results at the stimuli level, and inferences can be made on the stimuli population. This makes the results replicable with other sets of stimuli drawn from the same population, but only if they are administered to the same sample of respondents (Judd et al., 2012).

As an attempt to overcome the replicability and generalizability issues concerning either the *by-participant* approach or the *by-stimulus* approach, Raaijmakers, Schrijnemakers, and Gremmen (1999) suggested to report the results obtained with both the *by-participant* and the *by-stimulus* analyses. The results are then accepted as significant only if both analyses yielded significant results.

The underlying logic appears to be quite straightforward. Given that the *by-participant* analysis allows for generalizing to other populations of respondents (but only if the same stimuli are employed) and the *by-stimuli* analyses allows for generalizing to others popula-

tion of stimuli (but only if the same sample of respondents is used), then, if they are both significant, it is possible to generalize across both of them.

Quite blatantly, this approach cannot do what it claims to do (Raaijmakers et al., 1999; Raaijmakers, 2003). The *by-participant* analyses keep ignoring the sampling variation of the stimuli and the *by-stimuli* analyses keep ignoring the sampling variation of the respondents. As such, their results are still flawed by un-wanted and uncontrolled error variance. Besides, this approach presents also theoretical fallacies. The syllogistic reasoning according to which, if both the premises are true (both the *by-participant* and the *by-stimulus* analysis are simultaneously significant and the results can hence be replicated on different samples of respondents and stimuli, respectively), then also their conjunction would hold true (i.e., the results can be replicated with concurrently new samples of respondents and stimuli), appears too bold.

A solution to this impasse is to consider both the respondents and the stimuli as random factors. By doing so, all the sources of variability at the different levels, and their potential interactions, can be accounted for, resulting in more reliable estimates (Barr et al., 2013; Judd et al., 2012; Wolsiefer et al., 2017).

Considering the respondents and the stimuli as random factors implies that both levels are assumed to be drawn from larger populations, on which the researcher is interested in making inferences (Judd et al., 2012; Wolsiefer et al., 2017). While the implications for the respondents are immediately clear, since they are typically considered as drawn from larger population, the same cannot be said for the stimuli. Indeed, the stimuli are usually considered as fixed factors. They are hence supposed to represent their population, and that they do all have the same effect on the outcome measures.

Assuming that all employed stimuli in the IAT have the same effect on the outcome measure (i.e., the difference in the average response time in each associative condition) is an already proved fallacy (e.g., Bluemke & Friese, 2006; Ellithorpe et al., 2015). When stimuli are considered as just one the possible samples that can be drawn from a larger population, it is implicitly entailed that they can vary for each observational unit. Therefore, they are allowed to have a different functioning, or, in other words, that they can make a difference in the

observed measure, and that this difference and the information it conveys can be investigated (Judd et al., 2012).

The scores presented in Chapter 2 are all affected by the above mentioned issues, as well as the formal models for the analysis of the IAT data presented in Chapter 3. The approach used to address the fully-crossed structure of the IAT is presented in Chapter **??**.

1.4.1 More than one implicit measure

It is not uncommon to find studies in which the IAT and the SC-IAT are used concurrently to obtain both a comparative measure of the attitudes towards one object in comparison to its opposite, as well as an absolute measure of the positive/negative evaluations towards each of them (e.g., Bulmer & Izuma, 2018; Epifania, Anselmi, & Robusto, 2020a; Glashouwer, Vroling, de Jong, Lange, & de Keijser, 2013). To pursue this aim, one IAT and two SC-IATs, one for each of the target objects, are administered to the same respondents. The data of each implicit measure are separated and analyzed separately by computing individual scores for each measure. These scores are then employed for further analysis.

When both implicit measures are administered together, the fully-crossed structure represented in Table 1.3 is repeated for each implicit measure. In other words, each implicit measure comes with its own sources of variability due to its fully-crossed structure.

Moreover, a super-ordinate variable, above the associative conditions, is added. The new super-ordinate variable is the type of measure. The associative conditions are hence nested within the specific implicit measure, while the respondents, besides being crossed with the stimuli and the measure-specific associative conditions, are also crossed with the implicit measures. Moreover, since the same stimuli are usually employed to represent the target objects and the evaluative dimensions in all implicit measures, also the stimuli are crossed with the implicit measures.

However, not all the stimuli are crossed with all implicit measures. While the stimuli belonging to the evaluative dimensions are indeed crossed with all implicit measures (i.e., they are administered in all implicit measures), the stimuli representing the target objects are pre-

sented only according to the specific measure. Specifically, stimuli representing both target objects are presented in the IAT, while only one of the target object categories is presented in each SC-IAT. The variable type of measure is hence introducing a nesting related to the stimuli.

The implicit measure (e.g., the IAT rather than the SC-IAT) tends to be a variable of interest only in those studies aimed at either the validation of the measure itself or at the investigation of their different functioning. Nevertheless, the by-measure variability that has been introduced needs to be taken into account to obtain reliable estimates and scores.

The scores computed on each implicit measure hence include the sources random variation due to both the fully-crossed design of each measure and the variability that should be expected by the multiple administration of the same set of stimuli to the same sample of respondents. The approach aimed at overcoming these issues with a comprehensive modeling of multiple implicit measures is presented in Chapter 6.

Chapter 2

Typical scoring of implicit measures

This chapter presents typical scoring procedures for the IAT and the SC-IAT data, as well as the development of new, open source tools for easily computing the scores for the IAT and the SC-IAT.

The typical scoring algorithms for both the IAT and the SC-IAT are illustrated and described. It is not unusual to find studies in which both the IAT and the SC-IAT have been administered together, and their performances, for example for predicting a behavioral outcome, have been compared. However, this comparison might be biased by many differences concerning both the administration and the scoring procedures of the two implicit measures. Therefore, new scoring algorithms have been introduced with the aim of reducing the noise due to external factors (i.e., the scoring procedure itself) in the comparison between implicit measures. The results of an empirical study in which the performance of typical and modified scoring algorithms have been compared in respect to the prediction of a behavioral outcome are reported.

The core computation of both the IAT and the SC-IAT scores is rather easy. Nonetheless, the many steps that have to be undertaken for preparing and cleaning the data make it an error-prone procedure, and compromise the reproducibility of the results. Since there is a lack of easy-to-use and open source tools for their computation, a Shiny app and an `R` package have been developed for the computation of the IAT and the SC-IAT *D* scores. These tools are

presented at the end of this chapter.

2.1 The IAT *D* score

Greenwald et al. (2003) introduced different variations of the *D* score algorithm (Table 2.1). These variations result from the combination of the error correction strategies ("Error replacement" in the Table) and the treatment for fast responses ("Lower tail treatment" in the Table).

Table 2.1: Overview of *D* score algorithms.

Algorithm	Error replacement	Lower tail treatment
*D*1	Built-in correction	No
*D*2	Built-in correction	Delete trials $<$ 400ms
*D*3	Mean + 2*sd*	No
*D*4	Mean + 600ms	No
*D*5	Mean + 2*sd*	Delete trials $<$ 400ms
*D*6	Mean + 600ms	Delete trials $<$ 400ms

Note: For all algorithms, trials with latency $>$ 10,000ms are discarded. For the algorithm in which error responses are replaced with the average response time plus a penalty, the average response time is computed on correct responses only.

Blocks B1, B2, and B5 in Table 1.1 are considered as pure practice blocks and are hence discarded from the computation. Only trials from Blocks B3, B4 (i.e., Mapping A) and B6, B7 (i.e., Mapping B) are used for the computation.

The error correction strategies based on built-in correction (*D1* and *D2*) refer to the IAT procedure including feedback, according to which respondents have to correct their error responses. The response time considered for the computation of the *D* score is the response time at the first (incorrect) response inflated by the time required to correct it. All other

algorithms (from *D1* to *D6* in Table 2.1) use a *post-hoc* error correction strategy, for which error responses are replaced by the average response time of the correct responses in the block in which the error occurred increased by a standard penalty (i.e., either 600ms or twice the standard deviation). The other feature differentiating the *D* score algorithms is the lower tail treatment, according to which fast trials (trials faster than 400ms) are discarded or not.

Regardless of the specific features of each algorithm, the core procedure for computing the *D* score is the same. Firstly, the *D* scores of associative practice blocks (Eq. 2.1):

$$D_{\text{practice}} = \frac{M_{\text{B6}} - M_{\text{B3}}}{SD_{\text{B6, B3}}}, \tag{2.1}$$

and of associative test blocks (Eq. 2.2)

$$D_{\text{test}} = \frac{M_{\text{B6}} - M_{\text{B3}}}{SD_{\text{B6, B3}}}, \tag{2.2}$$

are computed. In both cases, the difference in the average response time between the two critical blocks is divided by the standard deviation computed on the pooled trials of both blocks. Once D_{practice} and D_{test} are obtained, it is possible to compute the actual *D* score:

$$D \text{ score} = \frac{D_{\text{practice}} + D_{\text{test}}}{2}. \tag{2.3}$$

The block order in Equation 2.1 and Equation 2.2 is arbitrary and can be reversed. The resulting *D* has to be interpreted accordingly. In the Coke-Pepsi IAT in Chapter 1, Blocks B3 and B4 constituted the Coke-Good/Pepsi-Bad condition. Conversely, the Pepsi-Good/Coke-Bad condition was composed of Blocks B6 and B7. If the *D* score is computed following the order of the blocks in Equation 2.1 (i.e., $M_{\text{B6}} - M_{\text{B3}}$) and in Equation 2.2 (i.e.,$M_{\text{B7}} - M_{\text{B4}}$), a positive score would indicate slower responses in Pepsi/Good-Coke/Bad condition than in Coke/Good-Pepsi/Bad one, probably indicating a preference for Coke over Pepsi. Vice versa, if the order of the Block in Equations 2.1 and 2.2 is reversed (i.e., $M_{\text{B3}} - M_{\text{B6}}$ and $M_{\text{B4}} - M_{\text{B7}}$, respectively), a positive score would indicate slower responses in the Coke/Good-Pepsi/Bad condition than in the Pepsi/Good-Coke/Bad condition, indicating a possible preference for

Pepsi over Coke.

2.2 The SC-IAT *D* score

Since blocks B1 and B3 (Table 1.2) are considered as pure practice blocks, they are discarded from the computation of the SC-IAT *D* score. If a rtw was included in the administration procedure, all responses exceeding it are considered as non-responses and are discarded from the computation. All responses with a latency faster than 350ms are discarded, and error responses are replaced with the average response time of the block in which the error occurred inflated by a standard penalty of 400ms.

After cleaning and preparing the data, the SC-IAT *D* score is simply computed as the difference in the average response time of the two critical blocks (i.e., $M_{\text{B4}} - M_{\text{B2}}$) divided by the standard deviation computed of the correct trials of both blocks. As for the IAT, the order of the critical blocks is arbitrary and the interpretation of the *D* score changes accordingly.

In the Coke SC-IAT example illustrated in Chapter 1, Block B2 was the Coke-Good condition, while Block B4 was the Coke-Bad condition. Following this structure, if the *D* score is computed by taking the difference between Blocks B4 and B2, a positive score would indicate slower responses in the Bad/Coke condition than in the Good/Coke one, standing for a positive evaluation of Coke. Vice versa, if the score is computed in the opposite direction (i.e., $M_{\text{B2}} - M_{\text{B4}}$), a positive score would indicate slower responses in the Good/Coke condition than in the Bad/Coke condition, indicating a plausible negative evaluation of Coke.

2.3 A fairer comparison between the IAT and the SC-IAT

In Study 1, Karpinski and Steinman (2006) directly investigated and compared the predictive ability of a Coke-Pepsi IAT, that of a Coke SC-IAT, and that of a Pepsi SC-IAT. the behavioral outcome was the choice between a can of Coke and a can of Pepsi. As their results suggested, measures obtained from both the Coke-Pepsi IAT and the Pepsi SC-IAT played a role in predicting the soda choice, while the measure obtained from the Coke SC-IAT did

not contribute to the choice prediction. Drawing on these results, authors speculated that the soda choice is more guided by a positive evaluation of Pepsi than by a negative evaluation of Coke. Nonetheless, the direct comparison between the predictive ability of the two implicit measures has been poorly investigated. Despite the study by Karpinski and Steinman (2006) provided interesting information on the functioning of implicit processes and the comparison between implicit measures, it also had some shortcomings that might have undermined the validity of their results. The aim of the study reported in this section was to to provide a fairer comparison of the predictive ability of the IAT and the SC-IAT in respect to a behavioral choice by presenting new scoring algorithms for the two implicit measures. This study is published in Epifania, Anselmi, and Robusto (2020a).

Among the shortcomings in Karpinski and Steinman (2006), the sample size was rather small, and results should hence be interpreted with caution. Moreover, the comparison between the predictive ability of the implicit measures might have been affected by issues concerning their administration and scoring procedures. The IAT and the SC-IAT differed in the number of trials, the number of blocks, and the number of exemplars representing each category. The SC-IAT employed more trials and more stimuli than the IAT, for both the evaluative dimensions (twenty-one exemplars for each SC-IAT evaluative dimension versus five exemplars for each IAT evaluative dimension) and the object categories (seven exemplars for each SC-IAT target object category and five exemplars for each IAT target object category). Furthermore, the administration of the SC-IAT included a rtw, while that of the IAT did not have such a constraint on the responses. The presence of a rtw makes the task more difficult and produces a sense of urgency that is otherwise missing (Karpinski & Steinman, 2006). Additionally, in the SC-IAT respondents were given feedback for each correct and incorrect response, while the IAT administration procedure did not include any feedback. The labels used for representing the positive and negative evaluative dimensions changed across implicit measures (*Pleasant* vs *Unpleasant* for the IAT and *Good* vs *Bad* for the SC-IAT), as well as the response keys used for sorting the stimuli. The IAT *D* score was computed according to the *D* score procedure in Greenwald et al. (2003), despite Karpinski and Steinman (2006) failed to report the exact algorithm they employed. The SC-IAT *D* score presented in Section

2.2 was used for computing the SC-IAT *D* score.

Given the differences between administration and scoring, the comparison between the ability of the IAT and that of the SC-IAT to predict a behavioral outcome might have been unfair. To the best of our knowledge, there is neither a scoring procedure employing the same criteria for both the IAT and the SC-IAT, nor an attempt to align the two implicit procedures to allow for a fairer comparison between their predictive ability. It would be interesting to compare the predictive ability of the two implicit measures by using the same scoring procedure on their data and by keeping the administration as similar as possible, while acknowledging their key features (e.g., block types and usual length of the blocks). If by using the same scoring procedure and by reducing the administration-related differences there are still differences in the predictive ability of the two measures, these differences can be reasonably attributed to the implicit procedure itself.

To obtain a fairer comparison between the two implicit measures, both administration (e.g., stimuli, rtw, feedback) and scoring of the two procedures have been aligned.

2.3.1 Method

To test the predictive ability of the new scoring procedures, one Chocolate IAT, one Dark chocolate SC-IAT, and one Milk chocolate SC-IAT were developed. The decision to use chocolate as the object category was driven by different reasons. Firstly, chocolate preference should not be sensitive to social desirability, and hence respondents would have no concerns in explicitly reporting their actual chocolate preference. Moreover, it offers the chance to ask for a behavioral choice disguised as a reward for the participation.

Inquisit 3.0 (Software, 2011) was used for administering the implicit measures (i.e., the IAT and the two SC-IATs) and the demographic questionnaire.

Participants. Participants were recruited at the University of Padova. One-hundred and sixty-one people (F $= 63.55\%$, Age $= 23.95 \pm 2.83$) volunteered to take part in the study, with no compensation. Participants were informed about the confidentiality of the data, and they were given the possibility to withdraw from the experiment at any time they wished.

They were asked for their consent to take part in the study. Majority of the participants were students (94.08%), including both undergraduates, master, and PhD students. Only two participants reported to have a PhD title, while the majority reported to have a bachelor's degree (43.42%), immediately followed by those who reported having a high school diploma (32.24%) and a master's degree (23.03%).

Materials and Procedure. Chocolate stimuli were composed by seven images of chocolate that were modified to represent either Dark or Milk chocolate. Seven images for each type of chocolate were used. Three independent judges evaluated the stimuli regarding their properties, specifically whether they were clearly identifiable as dark or milk chocolate images. The three judges agreed on the representativeness of the stimulus in respect to the category it was supposed to to belong. All chocolate images were presented on a white background.

In the Chocolate IAT, both Dark and Milk chocolate images were used. In the two SC-IATs, only either dark (Dark SC-IAT) or milk (Milk SC-IAT) chocolate images were used. Object categories were labeled *Dark* or *Milk*. Evaluative attributes categories were composed of 13 stimuli each. The evaluative categories were labeled as *Positive* (i.e., "good", "laughter", "pleasure", "glory", "peace", "happiness", "joy", "love", "wonderful", "beautiful", "excellent", "heaven", "marvelous") or *Negative* (i.e., "evil", "bad", "horrible", "terrible", "annoying", "pain", "failure", "hate", "nasty", "disaster", "agony", "ugly", "disgust"). Response key "E" was used for sorting the stimuli belonging to the categories represented on the left-side of the screen. Response key "I" was used for sorting the stimuli belonging to the categories represented on the right side of the screen. The SC-IAT practice Blocks B1 and B3 were composed of 20 trials, as for the practice blocks of the IAT. Neither the IAT nor the SC-IATs included any feedback or rtw. Respondents were asked to be as fast and as accurate as they could in performing the tasks.

Respondents were explicitly asked to report their evaluation for Dark and Milk chocolate on two distinct items ("How much do you like Dark chocolate?" and "How much do you like Milk chocolate?") rated from "0 – Not at all" to "5 – Very much".The order of presentation of the implicit measures was counterbalanced across participants, while the demographic

questionnaire and the choice were kept constant at the end of the experiment.

As a reward for their participation, respondents were offered with either a free Dark chocolate bar or a Milk chocolate one. The experimenter registered respondents' choices after they left the laboratory.

Chocolate IAT: The critical blocks were composed of 60 trials each (20 practice and 40 test), defining the Dark-Good/Milk-Bad condition (DGMB), and the Milk-Good/Dark-Bad condition (MGDB).

Dark SC-IAT: The critical blocks were composed of 72 trials each, defining the Dark-Good/Bad (DG) and the Good/Dark-Bad (DB) conditions.

Milk SC-IAT: As for the Dark chocolate SC-IAT, the critical blocks were composed of 72 trials each, defining the Milk-Good/Bad (MG) and the Good/Milk-Bad (MB) conditions.

2.3.2 Data analysis

Data cleaning and *D* score

All the IAT *D* score algorithms not including a built-in correction (algorithms *D*3, *D*4, *D*5 and *D*6 in Table 2.1) were computed. The procedure described in Section 2.2 was followed for computing the SC-IAT *D* score.

The scoring algorithms that have been introduced in this study result from different combinations of two main characteristics. One characteristic is the trials on which the standard deviation for the replacement of error responses was computed (i.e., only correct trials or all trials). The other characteristic is the quantity used for standardizing the difference in the response times between the associative conditions (i.e., Cohen's pooled standard deviation or pooled trials standard deviation). Cohen's pooled standard deviation for two groups of size n_1 and n_2 is as follow:

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1 + (n_2 - 1)SD_2}{n_1 + n_2 - 2}} \tag{2.4}$$

The resulting eight combinations (identified by letter "*m*", *modified*) are illustrated in

Table 2.2.

Table 2.2: Overview of modified algorithms for computing the IAT and the SC-IAT scores.

Feature	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$	$m7$	$m8$
Lower tail treatment	< 350ms							
Upper tail treatment	$> 10{,}000$ms							
Error treatment	Mean (Correct) + 2 *sd* (Correct)				Mean (Correct) + 2 *sd*			
Denominator	Pooled trials		Cohen		Pooled trials		Cohen	
Denominator trials	Correct	All	Correct	All	Correct	All	Correct	All

While the typical procedure for the SC-IAT includes a default lower tail treatment, the lower tail treatment for the IAT depends on the specific D score algorithm (see Table 2.1). To have a comparable score, a common lower tail treatment for both procedures was set (i.e., responses with a latency less than 350ms were discarded). Since it is not uncommon to find SC-IATs with no rtw, a common upper tail treatment for response times was proposed for both implicit measures (i.e., responses over 10,000ms were discarded). Concerning the SC-IAT upper tail treatment, it might be argued that the deletion of the responses higher than 1,500ms (i.e., the rtw cut-off) would be a more appropriate threshold for slow responses. Nonetheless, the presence of the rtw itself produce an urge to respond that is missing when the rtw is not included in the administration procedure (Karpinski & Steinman, 2006).

Since the SC-IAT is known to be an easier task than the IAT (Karpinski & Steinman, 2006), the latency of the responses in the SC-IAT tend to be faster than the latency of the responses in the IAT. Therefore, assuming 600ms as a reasonable time for correcting the error response might be a too strong assumption for the SC-IAT data. Conversely, the penalty used in the SC-IAT (400ms) might be not enough for acknowledging the response time needed for correcting the error response in the IAT. For this reason, the error responses are replaced by the average response times in the block in which the error occurred inflated by two times the standard deviation of the block.

The pooled trials standard deviation and the Cohen's pooled standard deviation were computed either considering only correct responses or all trials. In the former case, the variability due to incorrect responses is not accounted for, while it is addressed in the latter case. Finally, the IAT modified procedures were computed as the difference between the two associative conditions, instead of as the mean of the standardized average response time differences between the practice and test blocks.

The IAT scores (typical and modified) were computed so that positive scores indicated faster responses in associating Milk chocolate with positive attributes and Dark chocolate with negative attributes, and hence an implicit preference for Milk chocolate over Dark chocolate. Conversely, negative scores indicated faster responses in associating Dark chocolate with positive attributes and Milk chocolate with negative attributes, indicating an implicit preference for Dark chocolate over Milk chocolate.

For the SC-IATs, both typical and modified procedures were computed so that positive scores indicated faster responses in associating the target chocolate with positive attributes than with negative attributes, hence indicating an implicit positive evaluation of the target chocolate. Conversely, negative scores indicated faster responses when the target chocolate was associated with negative attributes, and an implicit negative evaluation of the target chocolate.

Consistency between modified and typical scores, and relationship with explicit measures

Pearson's correlations between explicit chocolate evaluations, typical and modified scoring were computed. Moreover, Pearson's correlations were computed between the typical and modified scores to check for their consistency.

Prediction of the behavioral outcome

The typical and the modified scores are regressed on the chocolate choice, coded as 0 for the Dark chocolate choice (DCC) and 1 for the Milk chocolate choice (MCC). Each score is

regressed on the choice in a separate logistic regression. Since the choice is presented as a dichotomous task in which Dark chocolate is contrasted with Milk chocolate, it is plausible that the relative preference for one chocolate over the other plays a role in determining the actual choice. The score of each SC-IAT conveys a unique information on the absolute positive or negative evaluation of one type of chocolate. As such, each of them lacks a part of information that might be crucial in predicting the choice. The use of the linear combination of both SC-IATs scores or a combined score might solve this issue. However, since the SC-IATs scores are obtained from two different experiments, their combination, either linear or in a comprehensive score, might be considered a stretch. Grounding on these considerations, both the single Dark SC-IAT score and the single Milk SC-IAT score, as well as their linear combinations, were used for predicting the choice.

Nagelkerke's R^2 (Nagelkerke, 1991) and model accuracy of prediction (Faraway, 2016) are used as criteria for investigating the scores best accounting for the behavioral choice. Specifically, model general accuracy (i.e., ratio between the number of chocolate choices correctly identified by the model and the total number of choices), DCC accuracy (i.e., ratio between the number of DCCs correctly identified by the model and the total number of observed DCCs), and MCC accuracy (i.e., ratio between the number of MCCs correctly identified by the model and the total number of observed MCCs) are computed.

2.3.3 Results

Data from nine participants were discarded. Eight of them explicitly reported not understanding the tasks they were asked to perform in either the IAT or one of the SC-IATs, while one of them registered too many fast responses, specifically on the Dark chocolate SC-IAT (more than 30% of responses with a latency lower than 350ms). The final sample was composed of 152 participants (F $= 63.82\%$, Age $= 24.03 \pm 2.82$). Milk chocolate was chosen by the 48.03% of the participants.

The median for the explicit evaluation of Dark chocolate was 3 (Q_1 = 2, Q_3 = 5). The median for the explicit evaluation of Milk chocolate was 4 (Q_1 = 3, Q_3 = 4). No trials

exceeding the threshold of 10,000ms were found in the SC-IATs. Three trials exceeding the 10,000ms threshold were found in the IAT, and they were eliminated. The lowest percentages of trials faster than both 400ms (1.39%) and 350ms (0.19 %) were found in the IAT. The two SC-IATs showed similar percentages of trials faster that 350ms (1.00% and 0.90% in Milk SC-IAT and Dark SC-IAT, respectively), as well as of trials faster than 400ms (4.40% and 4.32% in Dark SC-IAT and in Milk SC-IAT, respectively). All implicit measures had the same overall percentage of correct responses (95%).

In the IAT, the overall average response time was 862.03ms (*sd* $=$ 496.50, *skewness* $= 3.45$, *kurtosis* $= 22.01$). The average response time in the DGMB condition was 976.44ms (*sd* $=$ 555.19, *skewness* $=$ 2.88, *kurtosis* $=$ 14.01) and that in the MGDB condition was 747.62ms (*sd* $= 398.30$, *skewness* $= 4.76$, *kurtosis* $= 48.62$).

The overall average response time in the Dark SC-IAT was 679.45ms (*sd* $=$ 328.72, *skewness* $=$ 4.10, *kurtosis* $=$ 27.94). The average response time in the DB condition was 673.71ms (*sd* $=$ 322.87, *skewness* $=$ 3.90, *kurtosis* $=$ 24.46) and that in the DG condition was 685.19ms (*sd* $= 334.39$, *skewness* $= 4.27$, *kurtosis* $= 30.86$).

The overall average response time in the Milk SC-IAT was 675.90ms (*sd* $=$ 322.31, *skewness* $=$ 4.48, *kurtosis* $=$ 38.19). The average response time in the MB condition was 695.72ms (*sd* $=$ 344.84, *skewness* $=$ 4.10, *kurtosis* $=$ 28.32) and that in the MG condition was 656.08ms (*sd* $= 296.78$, *skewness* $= 4.98$, *kurtosis* $= 54.05$).

Relationship with explicit measures. Descriptive statistics for the typical scoring of all implicit measures, along with their correlation with explicit measures, are reported in Table 2.3.

Table 2.3: Descriptive statistics of the scores and correlations (r) with explicit chocolate evaluations.

	Modified	*M (sd)*	*Min*	*Max*	r_{Milk}	r_{Dark}	Typical	*M (sd)*	*Min*	*Max*	r_{Milk}	r_{Dark}
IAT	*m1*	$0.64\,(0.62)$	-1.91	1.72	0.40^{***}	-0.36^{***}	*D3*	$0.41\,(0.41)$	-1.29	1.25	0.42^{***}	-0.38^{***}
	m2	$0.64\,(0.60)$	-1.86	1.69	0.40^{***}	-0.37^{***}	*D4*	$0.39\,(0.39)$	-1.26	1.27	0.41^{***}	-0.38^{***}
	m3	$0.73\,(0.73)$	-2.25	2.84	0.39^{***}	-0.34^{***}	*D5*	$0.40\,(0.41)$	-1.29	1.29	0.42^{***}	-0.37^{***}
	m4	$0.72\,(0.70)$	-2.12	2.59	0.39^{***}	-0.35^{***}	*D6*	$0.39\,(0.39)$	-1.26	1.32	0.41^{***}	-0.37^{***}
	m5	$0.64\,(0.63)$	-2.29	1.72	0.40^{***}	-0.35^{***}						
	m6	$0.64\,(0.60)$	-1.85	1.69	0.40^{***}	-0.36^{***}						
	m7	$0.72\,(0.75)$	-2.34	2.85	0.39^{***}	-0.34^{***}						
	m8	$0.72\,(0.71)$	-2.11	2.60	0.39^{***}	-0.35^{***}						
Dark SC-IAT	*m1*	$-0.06\,(0.35)$	-0.98	1.07	-0.22^{**}	0.18^{*}	*D-Dark*	$-0.05\,(0.31)$	-0.74	0.78	-0.19^{*}	0.17^{*}
	m2	$-0.06\,(0.34)$	-1.03	0.94	-0.23^{**}	0.17^{*}						
	m3	$-0.06\,(0.36)$	-1.01	1.07	-0.22^{**}	0.17^{*}						
	m4	$-0.06\,(0.35)$	-1.05	0.94	-0.22^{**}	0.17^{*}						
	m5	$-0.06\,(0.36)$	-1.00	1.13	-0.20^{*}	0.16^{*}						
	m6	$-0.06\,(0.35)$	-0.95	0.99	-0.20^{*}	0.16						
	m7	$-0.06\,(0.36)$	-1.04	1.13	-0.19^{*}	0.16						
	m8	$-0.06\,(0.35)$	-0.97	1.00	-0.20^{*}	0.16						
Milk SC-IAT	*m1*	$0.16\,(0.39)$	-1.92	1.22	0.17^{*}	0.04	*D-Milk*	$0.15\,(0.33)$	-0.93	1.21	0.13	0.06
	m2	$0.16\,(0.39)$	-1.93	1.13	0.17^{*}	0.04						
	m3	$0.16\,(0.41)$	-1.92	1.5	0.17^{*}	0.04						
	m4	$0.16\,(0.40)$	-1.94	1.38	0.17^{*}	0.04						
	m5	$0.16\,(0.38)$	-1.39	1.23	0.15	0.05						
	m6	$0.16\,(0.37)$	-1.40	1.14	0.15	0.05						
	m7	$0.17\,(0.39)$	-1.39	1.51	0.16^{*}	0.05						
	m8	$0.16\,(0.39)$	-1.40	1.39	0.16^{*}	0.05						

Note: $^{***}\,p < .001$, $^{**}\,p < .01$, $^{*}\,p < .05$. *m*: Modified scoring algorithms; *D*: Typical scoring algorithms

Regardless of the scoring algorithm, the SC-IAT scores tended to have smaller effect sizes than the IAT scores. The IAT modified scores showed higher effect sizes than the IAT typical scores. The modified and typical SC-IAT scores were more consistent between each other.

The explicit Dark chocolate evaluation negatively and moderately correlated with the explicit Milk chocolate evaluation ($r = -.39$, $p < .001$). The IAT and the Dark SC-IAT typical scores significantly correlated with both explicit chocolate evaluations. The Milk SC-IAT typical score correlated with neither the Dark nor the Milk explicit evaluation. The IAT modified scores significantly and moderately correlated with the explicit evaluations of both Dark and Milk chocolate. The modified scores of both SC-IATs significantly correlated with the explicit evaluation of Milk chocolate. Only the first four modified scores of the Dark SC-IAT significantly correlated with the explicit evaluation of Dark chocolate. The correlation between the explicit evaluation of Dark chocolate and the scores of the Milk SC-IAT (both typical and modified) was near zero.

Consistency between typical scores, modified scores, and explicit measures. The correlation coefficients between the typical IAT *D* scores ranged between .99 and 1.00 (all ps $< .001$). The correlations between the typical IAT scores and the Dark SC-IAT typical score were all $-.21$ (all ps $< .01$). No correlations were found between the typical IAT *D* scores and the typical Milk SC-IAT score (correlations ranged between $-.04$ and $-.03$, all ps $> .05$). The typical *D-Dark* and *D-Milk* positively correlated between each other ($r = .15$, $p > .05$), but the correlation was not significant. The correlations between IAT modified scores ranged between .97 and .99 (all ps $< .001$). Their correlations with modified the *D-Dark* ranged between $-.31$ and $-.28$ (all ps $< .001$). The modified *D-Milk* and the modified IAT *D* scores did not correlate with each other (correlation coefficients ranged between $-.01$ and .01, all ps $> .050$). The correlations between modified *D-Dark* ranged between .98 and 1.00 (all ps $< .001$). The correlations between modified *D-Milk* ranged between .99 and 1.00 (all ps $< .001$). The correlations between modified *D-Milk* and *D-Dark* showed the same direction as the correlation between typical SC-IAT scores, ranging between .15 and .20. Interestingly, the correlation between all modified *D-Milk* and modified *D-Dark* from 5 to 8 (i.e., the scores

in which error responses were replaced by the mean added with twice the standard deviation computed on all trials) showed slightly stronger and significant correlations, ranging from .18 and .20 (all ps $< .010$).

Behavioral outcome. Results of the logistic regressions for predicting the chocolate choice are reported in Table 2.4.

Table 2.4: Choice prediction results, single predictors..

	$B\,(se)$	*Pseudo* R^2	*General*	*DCC*	*MCC*		$B\,(se)$	*Pseudo* R^2	*General*	*DCC*	*MCC*
IAT *m1*	$1.35^{***}\,(0.34)$	0.16	0.64	0.65	0.63	*D3*	$2.23^{***}\,(0.54)$	0.18	0.64	0.63	0.66
IAT *m2*	$1.38^{***}\,(0.35)$	0.16	0.62	0.62	0.63	*D4*	$2.26^{***}\,(0.55)$	0.18	0.64	0.66	0.63
IAT *m3*	$1.07^{***}\,(0.28)$	0.15	0.63	0.65	0.62	*D5*	$2.18^{***}\,(0.53)$	0.18	0.63	0.63	0.63
IAT *m4*	$1.12^{***}\,(0.29)$	0.15	0.64	0.66	0.63	*D6*	$2.22^{***}\,(0.55)$	0.17	0.64	0.65	0.63
IAT *m5*	$1.35^{***}\,(0.34)$	0.16	0.64	0.65	0.63						
IAT *m6*	$1.38^{***}\,(0.35)$	0.16	0.64	0.63	0.64						
IAT *m7*	$1.07\,(0.28)$	0.15	0.63	0.65	0.62						
IAT *m8*	$1.12\,(0.29)$	0.15	0.64	0.66	0.63						
Dark *m1*	$-0.73\,(0.48)$	0.02	0.53	0.62	0.44	*D-Dark*	$-0.70^{***}\,(0.54)$	0.01	0.53	0.65	0.41
Dark *m2*	$-0.72\,(0.49)$	0.02	0.53	0.62	0.42						
Dark *m3*	$-0.70\,(0.47)$	0.02	0.53	0.62	0.44						
Dark *m4*	$-0.69\,(0.48)$	0.02	0.52	0.62	0.41						
Dark *m5*	$-0.62\,(0.47)$	0.02	0.52	0.63	0.4						
Dark *m6*	$-0.63\,(0.48)$	0.02	0.52	0.63	0.4						
Dark *m7*	$-0.60\,(0.46)$	0.02	0.51	0.63	0.38						
Dark *m8*	$-0.60\,(0.47)$	0.01	0.51	0.63	0.38						
Milk *m1*	$0.33\,(0.42)$	0.01	0.53	0.77	0.26	*D-Milk*	$0.35^{***}\,(0.49)$	0.01	0.53	0.78	0.26
Milk *m2*	$0.33\,(0.43)$	0.01	0.53	0.77	0.26						
Milk *m3*	$0.32\,(0.40)$	0.01	0.52	0.76	0.26						
Milk *m4*	$0.32\,(0.41)$	0.01	0.53	0.77	0.26						
Milk *m5*	$0.31\,(0.44)$	0	0.5	0.75	0.23						
Milk *m6*	$0.31\,(0.44)$	0	0.52	0.76	0.26						
Milk *m7*	$0.30\,(0.42)$	0	0.5	0.75	0.23						
Milk *m8*	$0.30\,(0.42)$	0	0.51	0.76	0.25						

Note: $^{***}\ p < .001$. Bs are the log-odds for the probability of choosing Milk chocolate; *m*: Modified scoring algorithms; *D*: Typical scoring algorithms, Gen: General accuracy of prediction, DCC: Dark Chocolate Choice accuracy of prediction, MCC: Milk Chocolate Choice accuracy of prediction.

The IAT scores outperformed the scores of both SC-IATs in predicting the chocolate choice. The models including the IAT scores showed the highest values of Nagelkerke's R^2, and they resulted in a better accuracy of the prediction of both types of chocolate. Both SC-IATs scores showed low values of Nagelkerke's R^2, particularly Milk SC-IAT ones.

Typical and modified IAT scores tended to have similar values of both Nagelkerke's R^2 and accuracy of prediction. All the modified scores of the Dark SC-IAT resulted in slightly higher values of Nagelkerke's R^2. Only the first four modified Milk chocolate SC-IAT scores showed slightly higher Nagelkerke's R^2 than the typical ones. Modified SC-IATs scores showed a slightly worse performance than the typical ones.

The results of the choice prediction provided by the linear combination of the scores of each SC-IAT are reported in Table 2.5.

Table 2.5: Choice prediction results: SC-IAT scores linear combination.

	B_{Dark} *(se)*	B_{Milk} *(se)*	*Pseudo* R^2	Gen	DCC	MCC
D-Dark + *D-Milk*	$-0.77\,(0.55)$	$0.46\ (0.50)$	0.02	0.55	0.67	0.41
$m1_{\text{Dark}}$+ $m1_{\text{Milk}}$	$-0.81\,(0.49)$	$0.44\,(0.43)$	0.03	0.56	0.66	0.45
$m2_{\text{Dark}}$+ $m2_{\text{Milk}}$	$-0.80\,(0.50)$	$0.44\,(0.44)$	0.03	0.53	0.66	0.40
$m3_{\text{Dark}}$+ $m3_{\text{Milk}}$	$-0.78\,(0.48)$	$0.43\,(0.41)$	0.03	0.54	0.66	0.41
$m4_{\text{Dark}}$+ $m4_{\text{Milk}}$	$-0.77\,(0.49)$	$0.43\,(0.42)$	0.03	0.54	0.67	0.40
$m5_{\text{Dark}}$+ $m5_{\text{Milk}}$	$-0.71\,(0.48)$	$0.44\,(0.45)$	0.02	0.55	0.68	0.40
$m6_{\text{Dark}}$+ $m6_{\text{Milk}}$	$-0.72\,(0.49)$	$0.44\,(0.45)$	0.02	0.54	0.68	0.38
$m7_{\text{Dark}}$+ $m7_{\text{Milk}}$	$-0.68\,(0.47)$	$0.42\,(0.43)$	0.02	0.55	0.68	0.40
$m8_{\text{Dark}}$+ $m8_{\text{Milk}}$	$-0.69\,(0.48)$	$0.42\,(0.44)$	0.02	0.54	0.68	0.38

Note: Bs are the log-odds for the probability of choosing Milk chocolate. *m*: Modified scoring algorithms; *D*: Typical scoring algorithms; Gen: General accuracy of prediction, Pseudo *Pseudo* R^2: Nagelkerke's R; DCC: Dark Chocolate Choice accuracy of prediction; MCC: Milk Chocolate Choice accuracy of prediction.

The linear combination of the SC-IATs scores resulted in a better prediction of the chocolate choice than that provided by their singular scores. Their performance was still outperformed by that of the IAT. The coefficients of the typical and the modified *D-Dark* scores

tended to be higher than the coefficients of the typical and the modified *D-Milk* scores. The linear combination of the first four scores resulted in higher Nagelkerke's R^2 values, both in comparison with the typical scores and with the last four modified scores. The general accuracy and the DCC accuracy were similar across all scores, while the MCC accuracy showed a higher variability. Specifically, the linear combination of modified scores *m6* and that of modified scores *m8* showed the worst performance of all. The linear combination of *m1* resulted in the highest MCC accuracy.

As a final analysis, the incremental validity of the IAT and the two SC-IATs in respect to the self-report chocolate evaluations was investigated. Four hierarchical multiple logistic regressions for predicting the chocolate choice were specified for each of the scoring procedure. In the first step, the explicit evaluation of Dark and that of Milk chocolate were included. The IAT *D* scores entered at the second step. The *D-Dark* entered at the third step, and the *D-Milk* entered at the fourth step. This procedure was followed for both typical and modified scores. Nagelkerke's R^2 was used as a criterion to decide whether the added predictor was useful to account for the chocolate choice. Nagelkerke's R^2 at the first step (i.e., the model including only the explicit chocolate evaluations) was 0.83. From the second step on, the Nagelkerke's R^2 remains 0.84 for both typical and modified scores. It is reasonable to argue that the scores of implicit measures do not add anything to the prediction given by the explicit measures. However, this result should be interpreted with caution because the explicit chocolate evaluation was asked right before the behavioral choice.

2.3.4 Final remarks

By aligning the administration of the IAT and the SC-IAT, as well as their scoring methods, as much as possible it was possible to fairly investigate their relationship with explicit measures and their ability to predict behavioral outcomes. Consistently with the assumptions underlying the functioning of the two measures, the IAT scores highly correlated with both explicit chocolate evaluations, while the scores of the SC-IAT tended to correlate with just one of the explicit chocolate evaluations.

The IAT outperformed both the SC-IATs in the prediction of the behavioral choice. Taking these considerations together, it is possible to argue that the IAT has a better predictive capacity than the SC-IAT.

However, the higher predictive ability of the IAT scores might be due to the characteristics of the choice task itself. Since participants were presented with two different bowls of chocolate bars and were invited to take just one of them, their liking and/or dislike for both types of chocolate were concurrently playing a role in determining the choice. A measure able to include the comparative evaluation of the chocolate types, like the IAT, might hence best account for the actual chocolate preference from which it derives a better predictive ability. Conversely, measures dealing with only one of the components of the chocolate evaluation, like the SC-IAT, might be disadvantaged. However, even when SC-IAT scores were considered concurrently the general accuracy of prediction was similar to the one obtained when the single scores were considered. This result support the claim according to which a measure accounting for the relative attitudes towards two contrasting objects results in a better prediction of the choice between alternative options.

Since the better performance of the IAT might have been due to both the choice task and the type of preference assessed, it would be interesting to compare the performance of the two implicit measures when a clearly contrasted category is not identifiable, such as in the self-esteem case. In such cases, the absolute measure provided by the SC-IAT should outperform the relative one provided by the IAT in predicting behavioral outcomes. Additionally, the SC-IAT might outperform the IAT in predicting behavioral outcomes when the choice task is not strictly dichotomous. For instance, respondents might be left free to choose between two chocolate bars, both of them, or none of them, or even between different types of candy bars, including Dark and Milk chocolate ones.

Both in Karpinski and Steinman (2006) and in this study, the predictive validity of implicit measures was assessed for non-socially relevant stimuli, like soda and chocolate preference. Future studies should investigate the IAT and SC-IAT predictive validity in respect to socially relevant stimuli, such as members of stigmatized social groups. In pursuing this aim, different behavioral indicators might be used as a dependent variable, such as the willingness to

affiliate with members of the stigmatized social group. Another one could be having or not contacts with members of the stigmatized social groups.

Finally, Karpinski and Steinman (2006) administered the SC-IAT with a rtw at 1,500ms, while in this study it was not used. We could have applied an a posterior threshold for upper tail responses as if a rtw had been used in the administration. However, we decided not to do so because using a rtw affects respondents' performance. Indeed, Karpinski and Steinman (2006) observed that the presentation of the rtw produced a sense of urgency for giving the response that was missing when the rtw was not included. Future research on the systematic comparison between the two implicit measures might include a response time window for both the IAT and the SC-IAT.

2.4 R development

2.4.1 DscoreApp

Different options are available for computing the IAT *D* score, including SPSS syntax, Inquisit scripts, and `R` packages.

Inquisit scripts are the most straightforward way for obtaining the *D* score, since they compute it right after the IAT administration procedure, store the results along with other information on participants' performance (e.g., response time for each IAT trial, correct and incorrect responses), and do not require any programming skills. Nonetheless, these scripts work only when associated with the Inquisit administration procedure, can compute just one of the available *D* score algorithms at the time, and do not provide functions for visually inspecting the results. Finally, Inquisit requires a license to be used.

SPSS syntaxes provides several information on respondents' performance, and they are not tied to a specific administration software. Nonetheless, their use requires the SPSS license and a certain degree of expertise with SPSS language. They do allow for computing different *D* score algorithms by providing different scripts for the computation of each algorithm. No functions for directly plotting the results are provided.

The R packages (illustrated in Table 2.6) provides the open source alternative to both Inquisit and SPSS syntaxes. However, their use is not always straightforward because they require quite advanced programming skills, their functions are limited and they are often not well defined and understandable.

Table 2.6: Overview of the available packages for computing the IAT *D* score.

Package	Author	Functions	Multiple *D* score	Plot	Reliability
`IATanalytics`	Storage (2018a)	`IATanalytics`: Function to analyze raw data from an IAT `sampledata`: Sample dataset from a typical IAT	No	No	No
`IATScore`	Storage (2018b)	`BriefIAT`: Sample Brief IAT Data set (Abbreviated IAT) `IAT`: Sample IAT Dataset (Typical) `IATScore`: Score Implicit Association Test (IAT) output `TooFastIAT`: Sample IAT Dataset (Participant went too fast)	No	No	No
`IAT`	Martin (2016)	`cleanIAT`: Clean IAT data using the updated D-Scoring algorithm `IATData`: Sample Gender Stereotype Implicit Association Test data `plotIIV`: Plot intraindividual variability of reaction time `plotIndVar`: Plot individual variability in the IAT	Yes	Yes	Yes

		`plotItemErr`: Plot proportion of errors per item in the IAT `plotItemVar`: Plot IAT item variability			
`IATScores`	Costantini (2018)	`alg2param`: Convert the algorithm names to the generating parameters `Pretreatment`: Pretreat the IAT data in input `RobustScores`: Compute the Robust IAT scores `SplitHalf`: Split half reliability `TestRetest`: TestRetest `Tgraph`: Layout `qgraph` for multiple comparisons by package `nparcomp`	Yes	Yes	Yes

The `IATanalytics` and `IATscore` packages require for a specific arrangement of the columns of the data set to compute the *D* score, and they both include only the functions for computing the *D* score. The `IATscore` package includes a also a function for scoring the Brief IAT (B-IAT; Sriram & Greenwald, 2009). Both packages compute the *D* score for one respondent at a time, and no details on the specific algorithm are provided. The `IAT` package includes functions for cleaning the original data set, for plotting the data, and for computing different *D* score algorithms. However, it does not provide a clear labeling of each of the algorithms, but it allows the users to specify whether to discard the trials under 400ms or the error penalty to use. The combinations of these options result in the different *D* score algorithms. Additionally, the `IAT` package asks for a counter-intuitive coding of accuracy responses (i.e., 0 for correct responses, 1 for error responses). The plotting functions included in this package are not meant for plotting the *D* score results but the raw data.

The `IATScores` package appears to be the most complete package. Besides the functions for cleaning the data and for computing the typical IAT *D* score algorithms, it includes functions for the computation of the robust *D* score algorithms presented in Richetin, Costantini, Perugini, and Schönbrodt (2015). Additionally, it includes functions for computing the IAT reliability (i.e., test-retest reliability and split half). As for the `IAT` package, the plotting functions in the `IATScores` package are not meant for plotting the results of the *D* score computation.

DscoreApp (Epifania et al., 2019) was developed in `R` by means of packages `shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2018) and `shinyjs` (Attali, 2018) with the aim of providing an open source tool able to make the *D* score computation easier for researchers who commonly employ the IAT but have little or no programming experience. Furthermore, by providing an immediate representation of the results, it allows for a glimpse of the IAT results. DscoreApp can be retrieved at `http://fisppa.psy.unipd.it/DscoreApp/`. The source code of DscoreApp is available on GitHub (`https://github.com/OttaviaE/DscoreApp`). DscoreApp has been published in Epifania et al. (2019).

DscoreApp is organized in different panels (“Input”, “Read Me First”, “D-score results”, and “Descriptive statistics”). The setting options and the functions in the “Input” panel, as

well as the menu in the "Read Me First" panel, are interactive, so that users can easily access the information on DscoreApp functions and amenities.

The "Read Me First" panel provides important information on DscoreApp functioning, including an overview of the *D* score algorithms. A downloadable template suggested for using the app is provided (i.e., **Download template** button), even though it is not necessary to use it. DscoreApp is designed to work as long as the uploaded data set is in a CSV format, and includes the following variables: `participant` (i.e., participants' IDs), `latency` (i.e., latency of the responses in milliseconds), `correct` (i.e., accuracy of the responses, either 0 for error responses or 1 for correct responses), `block` (i.e., the labels identifying the four associative blocks of the IAT, B3, B4, and B6, B7). This panel also contains information on the downloadable file containing the results of the *D* score computation.

Users can either upload their own data set (i.e., by using the **Browse** button), or use the toy data set included in DscoreApp (i.e., by checking the `Race IAT dataset` checkbox) in "Input" panel. Once the data set is read, the app automatically populates the drop-down menus for choosing the labels denoting the four associative blocks, and the **Prepare data** button becomes clickable. When data are ready for the *D* score computation, the "Data are ready" message appears next to the **Prepare data** button, and all options for its computation and graphical display become active. Once a *D* score algorithm is chosen from the "Select your D" drop-down menu, the **Calculate & Update** button becomes active. Users can decide whether to eliminate participants whose error percentage exceeds a specified threshold (default is 25% according to Nosek, Banaji, & Greenwald, 2002) or whose fast responses (< 300ms) exceed 10% of the total responses (Greenwald et al., 2003). When these options are selected, participants exceeding the thresholds (if any) are not displayed in the "D-score results" panel. Every time a change in the configuration is made, the **Compute & Update** button must be clicked to apply the changes.

The "D-score results" panel (Figure 2.1) is populated once the **Calculate & Update** button is clicked for the first time. Both descriptive statistics of the results and their graphical representation are available at the same time, and they change interactively as users change the configuration in the "Input" panel. The `Summary` box reports the descriptive statistics

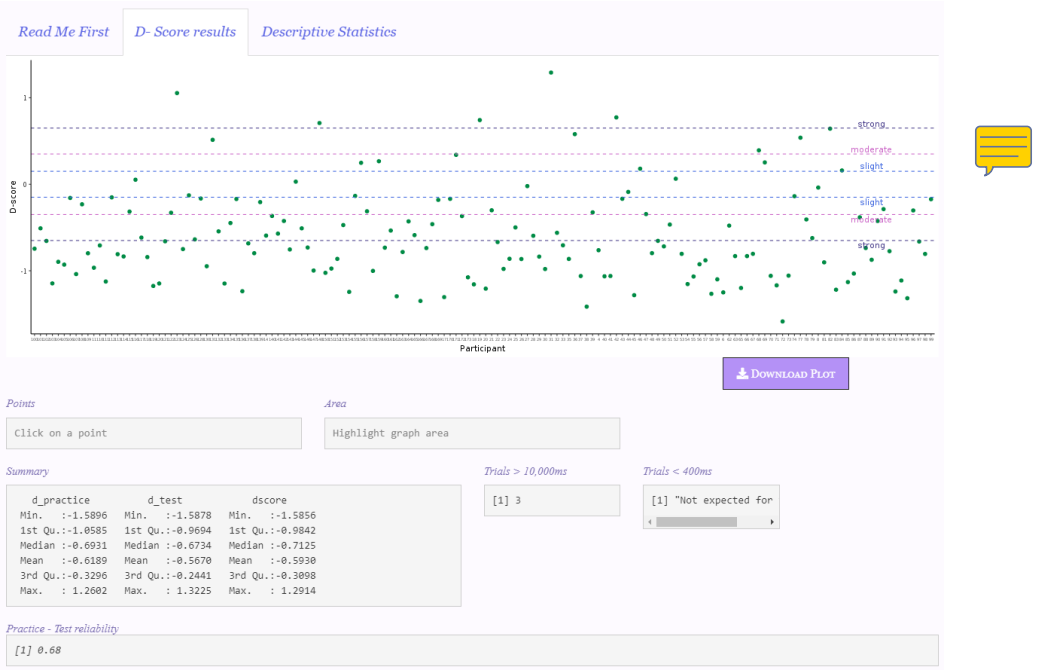


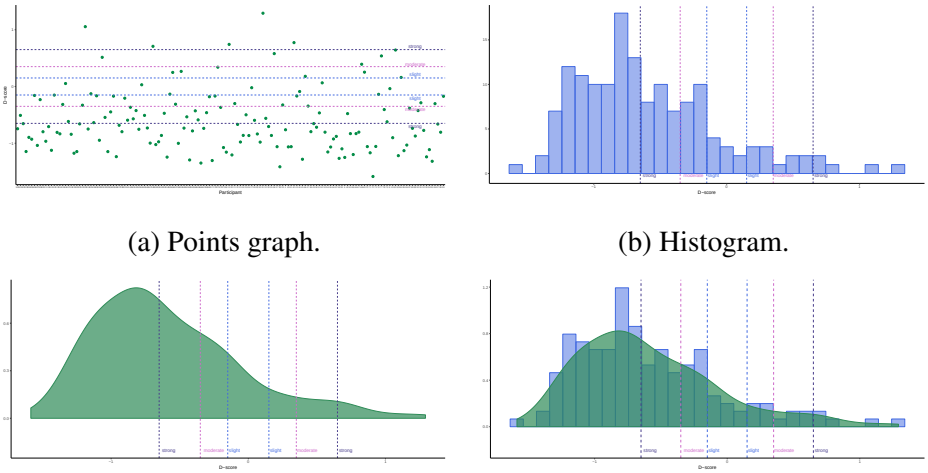
Figure 2.1: DscoreApp results panel.

of the D_{practice}, D_{test}, and the actual D score. The `Trials > 10,000ms` box reports the number of trials discarded because of a slow latency (if any), while the `Trials < 400ms` box reports the number of trials discarded because of fast response times. This box is populated only when a D score with fast trials deletion is selected, otherwise the "Not expected for this D" label is displayed. The `Practice-Test reliability` box contains the IAT reliability computed as the correlation between associative practice and associative test blocks across participants (Gawronski, Morrison, Phills, & Galdi, 2017).

Graphical representation is a convenient way to identify extreme scores or particular response patterns. Since it might be difficult to link a particular point (or points area) in the graph with the corresponding participants' IDs in the data set, DscoreApp comes with two handy tools designed to access the respondents' IDs from the graph. By clicking on a point in the graph, the ID of the participant that corresponds to the selected point, and his/her D score, appears in the `Points` box. By highlighting an area of the graph, the IDs of participants included in the area, along with their D scores appears in the `Area` box. The `Points` box and the *Area* box are represented in Figure 2.1, right underneath the graphical representation of

the results.

DscoreApp provides users with different options for the graphical representation of the results (Figure 2.2), at both individual (Figure 2.2a) and sample (Figure 2.2b, 2.2c, 2.2d) levels.



(a) Points graph.

(b) Histogram.

(c) Density.

(d) Histogram and Density.

Figure 2.2: Available graph representations.

All the graphical representations are downloadable in a .pdf format.

2.4.2 `implicitMeasures`

Despite both the IAT and the SC-IAT are commonly used for the implicit assessment of several constructs, `R` packages for the computation of only the IAT *D* score are available, while there are no packages for computing the SC-IAT *D* score.

Besides the above-mentioned shortcomings of `R` packages for computing the IAT *D* score, there are also issues related to the replicability of the results. Researchers' choice on which IAT *D* score algorithm to compute might influence the results and the conclusions that are drawn from them (Ellithorpe et al., 2015). Moreover, in many cases researchers fail to report of the exact *D* score algorithm they decided to use (Ellithorpe et al., 2015). Replicability issues might also rise from mistakes that might be caused by the many steps needed to clean and prepare the starting data set for computing the *D* score (Ellithorpe et al., 2015).

Despite DscoreApp addresses the majority of the replicability issues, it presents some drawbacks as well. Firstly, since the code is into the shiny interface, it cannot be called from the command line, making impossible to replicate it and hence the results. To be fair, this is a general and outstanding issue concerning shiny apps in general. This might not be a concern for general users, but it is indeed a problem in an open science framework where every code should be accessible and replicable with no effort and at any time. Moreover, the downloadable graphical representations are provided in a pdf format, and hence they cannot be further modified.

The `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020b, 2020c) was aimed at addressing the issues concerning both `R` packages and DscoreApp. It provides an easy and open source way to clean and score both the IAT and the SC-IAT, to easily compare different IAT *D* score algorithms, and to provide clear and customizable plots. Plot functions are all based on `ggplot2` (Wickham, 2016). The `implicitMeasures` package has been published in Epifania, Anselmi, and Robusto (2020c).

The source code of `implicitMeasures` is available on GitHub (`https://github.com/OttaviaE/implicitMeasures`). The package is downloadable from CRAN (`https://cran.r-project.org/web/packages/implicitMeasures/index.html`).

Table 2.7 provides an overview of the functions included in the `implicitMeasures` package.

The `implicitMeaures` package provides an easy way to compute the algorithms for the both the IAT and the SC-IAT in an automated way. By explicitly referring to the *D* score algorithm that has been used for computing the IAT score in the package, other users can easily replicate the results. Additionally, the possibility to compute all available algorithms for the IAT *D* score allows for an easy comparison between them. This makes possible to investigate whether or how the elimination of fast responses or the error replacement strategies affect the results.

All objects created with the functions in `implicitMeasures` functions can be ex-

Table 2.7: Contents and functions `implicitMeasures`.

Function	Description
`clean_iat()`	Prepare and clean IAT data
`clean_sciat()`	Prepare and clean SCIAT data
`compute_iat()`	Compute IAT *D* score
`compute_sciat()`	Compute SCIAT *D* score
`descript_d()`	Print descriptive table of *D* scores (also in LaTeX)
`d_density()`	Plot either IAT or SCIAT scores (distribution)
`d_point()`	Plot either IAT or SCIAT scores (points)
`IAT_rel()`	Compute IAT reliability
`multi_dsciat()`	Plot scores resulting from two SCIATs
`multi_dscore()`	Compute and plot multple IAT *D* scores
`raw_data()`	Example data set

ported in external files. For example, the data frame obtained from the `clean_iat()` function can be easily exported in a CSV file and then uploaded to DscoreApp (see Section 2.4.1).

The functions for plotting the results are based on `ggplot2` (Wickham, 2016), and they can be further modified by users, for instance by taking out the legend, adjusting the figure margins, changing labels and font. All the plots are then exportable as images (.jpg or .png) or as a .pdf.

Chapter 3

Formal modeling

This chapter provides a brief overview of the formal models introduced for modeling IAT data. These models are generally aimed at the investigation of the cognitive processes and the automatic associations involved during the performance at the IAT. Their advantages and drawbacks are outlined and discussed.

Some of these models are solely based on accuracy responses (Section 3.1), while others are able to concurrently model accuracy and time responses (Section 3.2). Despite the important and useful information provided at the sample level and/or the stimuli categories level, none of these models provides detailed information at the singular stimulus level.

The Rasch modeling of IAT data can overcome this issue, as illustrated in Section 3.3. Although this approach provides stimulus-specific information, it also comes with some drawbacks, mostly related to the discretization of response times and to the overlooking of the fully-crossed structure of the IAT.

3.1 Multinomial Models

3.1.1 The Quad Model

The Quad model (Conrey et al., 2005) is a multinomial processing tree model introduced for distinguishing the contribution of automatic processes from that of controlled processes in

driving respondents' performances at the IAT. The Quad model is entirely based on accuracy responses, and it exploits the logic of the assumption on which the IAT is based (i.e., response compatibility, according to which responses are faster and more accurate in the condition consistent with one's automatically activated associations).

According to this model, the observed accuracy responses are determined by the activation (or lack of thereof) of four qualitatively different processes, characterized by different levels of automaticity and controllability. These processes are the automatic activation of an association triggered by the target stimulus (*activation association*, AC), the ability to correctly identify the category to which the stimulus belongs (*discriminability*, D), the ability to overcome any automatically activated associations (*overcoming bias*, OB), and the influence of any response bias that may intervene in absence of any other process (*guessing*, G). A graphical representation of the Quad model is provided in Figure 3.1.

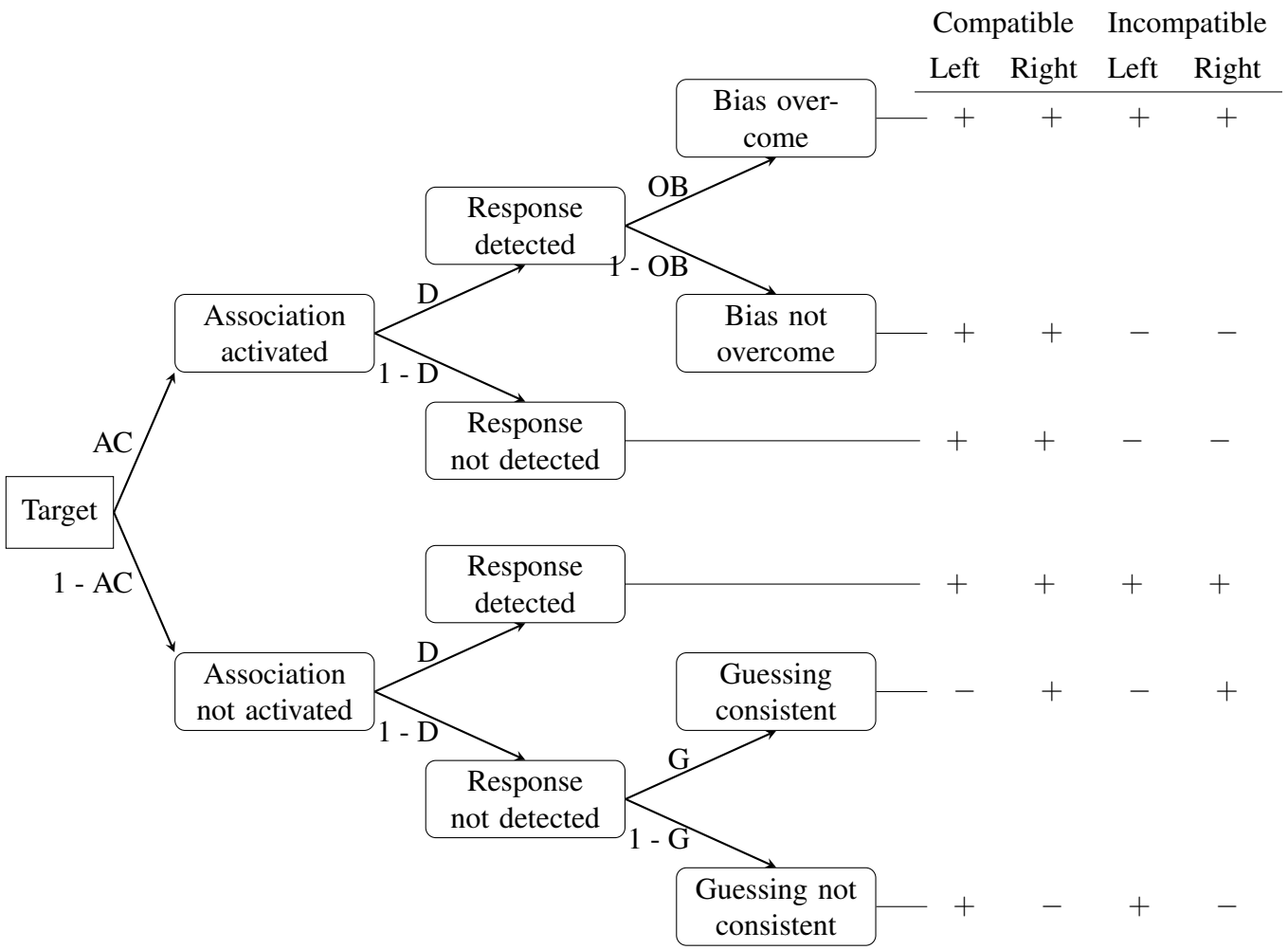


Figure 3.1: Quad Model (adapted from Conrey et al., 2005). Parameters with arrows pointing towards them are conditional on all the preceding ones. $+$: correct response (stimulus correctly assigned to its category), $-$: incorrect response (stimulus assigned to the incorrect category).

Each path in Figure 3.1 represents the likelihood that a parameter is activated. The activation of the parameters are conditional on the activation of the parameters preceding them.

The parameter AC describes the probability that an automatic association is activated by the triggering stimulus. This parameter expresses the most automatic component of the model and it is directly related to the strength of the association activated by the stimulus. The stronger the association between the stimulus and a negative/positive attribute, the more likely the activation of the automatic association.

The probability of giving the correct response does not solely depend on the activation of the automatic association, but also on the ability to identify the correct response among the available responses. This ability in turn depends on the availability of cognitive resources and on the application of some effort in determining the correct response. As such, parameter D represents the likelihood that the correct response *can* be identified, not the likelihood that the correct response *is* identified. The parameter D depends on several factors, including the motivation to have a good performance, the attention paid to the stimulus, and the availability of relevant information in memory.

If a negative automatic association is activated by the triggering stimulus, the respondent might try to fake the response in a socially desirable way. The parameter OB represents the likelihood that an activated bias is overcome in favor of a deliberate, and probably more desirable, response.

The processes described by the parameters D and OB are the ones that are mostly influenced by motivation and that mostly depend on the availability of cognitive resources to be activated and to drive the correct response. They do reflect two different aspects of controlled processes. The controlled process involved by the parameter D is an active search for the correct response, while the process described by the parameter OB exploits control for the inhibition of the response activated by the automatic association.

If none of the above-mentioned processes is activated, then the responses can be influenced by a response bias, such as the tendency to respond with the left response key. The parameter G represents the likelihood that a response bias, different than the automatic association, is activated and drives the responses.

As previously mentioned, the parameters of the Quad model are estimated from the observed proportions of a correct response given a stimulus type. Each arrow moving from left to right in Figure 3.1 represents the multiplication between the independent probabilities of each process. It results in the prediction of a specific response, either correct or incorrect. The sum of all probabilities associated to that response is the total probability of that response.

Consider a respondent with an implicit preference for Pepsi over Coke. For him/her, the incompatible condition is the Coke/Good-Pepsi/Bad condition. The probability that this respondent has of correctly sorting a can of Coke in the incompatible condition is given by the sum of the paths resulting in a correct response. In this case, three processes lead to the correct response. As such, the resulting equation is: P(correct|Coke, incompatible) $= AC \times D \times OB + (1-AC) \times D + (1-AC) \times (1-D) \times (1-G)$. The first path ($AC \times D \times OB$) represents the probability that an automatic association is activated by the stimulus (AC), that the response can be identified (D), and that the bias is successfully overcome (OB). The second path ($(1 - AC) \times D$) represents the probability that the automatic association is not activated ($1 - AC$) and that the response can be identified (D). Finally, the third path ($(1-AC) \times (1-D) \times (1-G)$) represents the probability that the association is not activated ($1-AC$), the correct response cannot be detected ($1-D$), and that automatically responding with the left response key is not an effect of guessing ($1 - G$). The sum of the products of the independent probabilities yielded by each path results in the total probability of a correct response to the stimulus.

By qualitatively disentangling the nature of the processes intervening during the performance at the IAT, the Quad model offers detailed information on the IAT functioning. Most importantly, the Quad model explicitly points out that the performance at the IAT should not be taken as the sole expression of automatic processes. Rather, the contribution of controlled processes should be acknowledged and taken into account for the explanation of social phenomena.

This point has crucial repercussion on applied researches using the IAT. For instance, in an IAT for the assessment of implicit prejudice, it would be of the uttermost importance to understand whether a resulting negative *D* score (e.g., a *D* score indicating preference for

White people over Black people) is actually due to the automatic activation of the association between one of the targeted groups and negative attributes or to other, more controlled, processes, such as the inability to identify the correct response. The information provided by the OB parameter allows for understanding whether a positive *D* score is the expression of genuinely automatic associations between the stigmatized group and positive attributes or by the desire to conceal negative attitudes towards the stigmatized group. Both the parameters AC and OB are associated with the typical IAT *D* score (Conrey et al., 2005). The *D* score was found to be positively associated with the parameter AC (i.e., the stronger the association activated by the stimulus, the higher the *D* score value). Moreover, the parameter AC allowed for pinpointing the contribution of distinct associations in a Race IAT, according to which both White-pleasant automatic associations and Black-unpleasant ones were related with the *D* score. The *D* score is hence capturing two distinct associations, and it is confounding them into a unique score. The *D* score was negatively associated to the ability of suppressing an automatic activated association described by the parameter OB. The higher the ability to overcome the bias, and hence the value of OB, the lower the IAT effect as expressed by the *D* score.

Grounding on these evidence, it can be said that the *D* score confounds different information into a single, generic score. Firstly, it is not possible to ascertain which of the specific automatic associations drives the performance at the IAT, leaving its meaning partially obscure. Moreover, controlled and automatic processes cannot be distinguished from one another, and their unique contribution is lost. It is not possible to ascertain whether the performance is driven by an actual automatic association or if the IAT effect reflects the respondents' ability to detect the correct response or their ability of overcoming an automatic activated bias. It appears evident that distinguishing between these processes is extremely important when inferences on sensitive psychological constructs, such as implicit bias, are made. Indeed, the implications of saying that a sample of individuals is implicitly biased towards a social group are extremely different from saying that the sample has a high ability in detecting the correct responses. The *D* score alone cannot be used as a measure of pure implicit bias.

The information provided by the Quad model are extremely useful and meaningful for a

correct interpretation of the IAT effect. However, it should be taken with caution for at least two reasons: The results are entirely based on accuracy responses and the persons' estimates are at the sample level and not at the individual respondent's level.

Regarding the first issue, the IAT is known to be an easy task – it is actually designed to be an easy task by choosing highly representative and easily sortable stimuli. As such, the error rates are extremely low, unless the respondent was distracted or the task itself did not work properly. This raises issues concerning the estimation of the model parameters and their reliability. Moreover, not using the time responses implies losing the majority of the information that can be retrieved from the IAT data, which can in turn lead to an incorrect interpretation of the results. For instance, the higher accuracy of the responses when the automatic associations is activated might be also associated to slower response times (speed-accuracy trade-off, Klauer et al., 2007). By considering only the accuracy responses, the Quad model is not able to rule out this possibility, and the conclusions based on the Quad model might be misleading. The parameter OB is the one controlling the accuracy of the responses when an automatic association is activated, and, as such, it should be the one mostly affected by the speed-accuracy trade-off. Indeed, results of Study 2 in Conrey et al. (2005) actually pointed in this direction. The parameter OB dropped significantly (i.e., respondents with automatically activated associations were not able anymore to provide the correct response) when a time constraint for giving the response was introduced. This result does indicate that the parameter OB captures a controlled process that needs time to be activated and successfully used. Not considering response times for interpreting the results on the parameter OB appears to be a fallacy leading to incorrect or misleading inferences.

Regarding the second issue, the estimates provided by the Quad model are at either the sample level or the stimuli categories level. Consequently, both the between–respondents variability and the between–stimuli one are completely ignored. Moreover, having sample estimates at the sample level for the respondents does not allow for investigating the respondents' individual differences, which is usually the main objective of applied social psychology. Similarly, the information at the level of the singular stimulus is neglected.

To be fair, in Study 4 in Conrey et al. (2005) respondent–specific estimates were obtained.

However, obtaining respondent–specific estimates with the Quad model is tricky given the high error rate needed in the starting contingency table, where stimuli categories are crossed with the associative conditions, within each respondent. To ensure an high error rate value for each possible combination, a longer IAT procedure should be adopted. As such, the Quad model is not feasible for investigating individual differences using IATs of typical length.

3.1.2 The ReAL Model

The ReAL model (Meissner & Rothermund, 2013) is a multinomial processing tree model based on accuracy responses. This model is aimed at mathematically distinguishing the contribution of the automatic associations from that of recoding or simplification strategies that might intervene during the performance at the IAT.

The ReAL model is based on two main assumptions. The first assumption is that only attitude objects activate automatic evaluative associations. Consequently, evaluative associations influence responding only for attitudes stimuli, while the same does not hold for attributes stimuli. In other words, attitude objects can be sorted according to their evaluative value, while evaluative attributes cannot be sorted according to the attitude objects to which they are associated. The second assumption logically follows from the first one. To be recoded into a unique category, target stimuli and evaluative attributes must be sharing a common feature, that is, an intrinsic positive or negative value. The common feature shared by the evaluative dimensions and the attitude objects is determined by attitude based associations. Attitude based associations can facilitate the correct sorting of the target stimuli in the condition consistent with individual's automatically activated association (i.e., compatible) but not in that against individual's own automatically activated association (i.e., incompatible). The recoding of target stimuli according to their evaluative dimension facilitates the performance only in the compatible condition, while it hinders it in the incompatible one.

For example, in the Coke-Good/Pepsi-Bad condition of a Coke-Pepsi IAT, a respondent with a strong preference for Coke might simplify the task by sorting Coke exemplars according to their positive value. As such, the task is reduced from a 4-choice task (i.e., *Coke* and

Good, *Pepsi* and *Bad*) to a 2-choice task (i.e., *Good*, which also includes *Coke*, and everything else). However, this strategy can work only in the associative condition that is consistent with respondent's automatically activated associations.

Grounding on these assumptions, three processes are assumed to drive the correct and incorrect response pattern in the IAT. Their influence on the performance changes according to the specific IAT associative condition, as illustrated in Figure 3.2. The illustration of the ReAL model is based on the Coke-Pepsi IAT example introduced in Chapter 1, considering a respondent with a strong preference for Coke over Pepsi.

Figure 3.2: ReAL model (adapted from Meissner & Rothermund, 2013). Parameters with arrows pointing towards them are conditional on all the preceding ones. +: correct response (stimulus correctly assigned to its category), −: incorrect response (stimulus assigned to the wrong category).

In the condition consistent with the respondent's automatically activated associations (i.e., compatible condition, top panel of Figure 3.2), three processes are assumed to play a potential role in driving the responses. If the stimulus appearing on the screen is combined with the evaluative dimension, the recoded category drives the response with a probability defined by the parameter *Re*. Since recoding is always associated with the correct response key, this process will always end in a correct response. When the recoded category is not activated, Label based processes (i.e., controlled search for the correct category to which the stimulus belongs and the associated response key) drive the response, with a probability defined by the parameter *L*. Also this process always ends in a correct response. The automatic evaluative associations (described by the parameter *A*) drive the response only when both the *Re* and *L* processes fail. If the association with *Good* drives the response, it will result in a correct response with a probability of *A*.

In the condition against respondent's own automatically activated associations (bottom panel of Figure 3.2), the recoding processes disappear. As in the compatible condition, when Label based processes fail, automatic associations drive the response. However, associations with *Good* lead to the incorrect response, while associations with *Bad* result in the correct response.

The structure of the model is identical for both target objects and evaluative dimensions. However, since evaluative dimensions cannot be activated by the attitude target objects, their association parameter is always fixed to .50.

The association parameter *A* estimated by the ReAL model offers some advantages over the parameter AC estimated by the Quad model. Firstly, in the ReAL model the association parameter is estimated separately for each target object category, while in the Quad model the automatic association parameter is estimated for the associated categories inferred from the associative condition. The separate estimates for each target object allow for investigating the nature (i.e., positive or negative) of the evaluative dimension activated by the target objects. Consequently, estimation of relative association strength can be avoided, overcoming one of the most criticized shortcomings of the measure derived from the IAT (e.g., Karpinski & Steinman, 2006). Nonetheless, caution should be used in the interpretation of separate

scores deriving from the IAT (Nosek et al., 2005). The Quad model does not include any parameter able to address potential recoding strategies. As such, the parameter *AC* might confound recoding strategies with the actual automatically activated associations (Meissner & Rothermund, 2013).

Recoding does contribute to the IAT effect by preventing task switching cost from attributes to target in the compatible condition, but not in the incompatible condition. The separate estimates of the evaluative associations activated by the target stimuli allow for a better understanding of the IAT effect. Indeed, the evaluative associations involved in the performance can be highlighted. For instance, the evaluative association estimates acknowledge the positive associations driving the IAT effect in a flowers-insects IAT as mostly determined by positive associations with *flowers* (Meissner & Rothermund, 2013).

The Label based processes described by the parameter *L* make possible to infer the easiness of the categorization task according to the stimuli categories. High values of *L* indicate that respondents often identified the correct responses grounding on the stimuli categories, as instructed by the task. The parameter *L* is sensitive to the type of stimuli presented. *L* results in higher values when the difficulty of the task is reduced (i.e., by presenting target images instead of words).

Consistently with the Quad model, the ReAL model considers the performance at the IAT as determined by both automatic and controlled processes. Both models are able to disentangle the automatic component from the automatic one by exploiting the information that can be retrieved from the accuracy responses. Differently from the Quad model, according to which the automatic associations are either immediately activated or not, the ReAL model posits the activation of the evaluative automatic associations only when the activation of all other processes fails. Moreover, the ReAL model considers the potential effect of the task-switching cost on the performance at the IAT by introducing a parameter (*Re*) for capturing controlled recoding strategies that can facilitate the performance. However, the ReAL model does not have a parameter explicitly describing the effort for overcoming the automatically activated bias.

The ReAL model presents some issues nonetheless. Firstly, it is not possible to rule out

the possibility that the parameter *Re* includes other non associative processes than just recoding, such as speed-accuracy trade-offs. The information provided by the parameter *L* does provide information about the difficulty of the categorization task according to the stimuli categories. However, this parameter is not able to disentangle the actual task difficulty from the respondent's ability to detect the correct response. Besides, it only provides an overall parameter for each stimuli category. The functioning of each stimulus is hence neglected. Moreover, as the Quad model, the ReAL model bases the estimation of its parameters on the error rates. To obtain an higher error rate than the ones usually obtained from typical IAT data, Meissner and Rothermund (2013) used IAT procedures with a response deadline. All respondents started with the same response deadline (i.e., 750ms), that was further adjusted grounding on their actual error rate. Specifically, it was shortened if the error rate was lower than 30%, and lengthened otherwise. Clearly, this makes the ReAL model applicable only to specific data set that are either already presenting an adequate error rate or that have been collected with the response deadline. Moreover, the difference in the administration procedures of the IATs in Meissner and Rothermund (2013) does not allow for a fair comparison between the predictive ability of the typical IAT scores and that of the model parameters obtained from the modified procedure (see Discussion of Study 7 in Meissner & Rothermund, 2013, for further details).

Finally, also the ReAL model provides only estimates at the sample level, so that its applicability for the investigation of individual differences is limited.

3.2 Time and Accuracy models

3.2.1 The Diffusion Model

The first application of the Diffusion Model (DM; Ratcliff, 1978) to IAT data can be found in Klauer et al. (2007). As the multinomial processing tree models presented in previous sections, the application of the DM to IAT data is aimed at disentangling the contribution of the process underlying the performance at the IAT. For pursuing this aim, the DM exploits all

information that can be retrieved from IAT data by mapping accuracy and time responses on the same metric.

The DM rests on the assumption that the decisions in two-choice tasks (such as the IAT) are based on processes of serial information accumulation over time. These processes begin from a starting point that lies between two thresholds. Each threshold is associated with a response. Once the information accumulation process reaches one of the thresholds, the response associated to that specific threshold is given. The average rate with which the information is accumulated over time (i.e., drift rate) does not always terminate at the same time, resulting in reaction time distributions, or at the same threshold, resulting in correct and error responses.

According to the DM, the decision-making process can be understood by considering four different components and their respective parameters: (i) the threshold separation (parameter a), (ii) the location of the starting point (parameter z), (iii) the average rate of information accumulation (i.e., drift rate, parameter v), and (iv) the nondecision component (t_0)

The amount of information that must be accumulated before one of the two responses is given is expressed by the parameter a. The parameter a assumes larger values when a high amount of information needs to be accumulated for giving the response, whereas it assumes smaller values when not much information is needed for the response. In the former case, the decision results in slower response times but higher accuracy, while in the latter one, it results in faster but less accurate responses. As such, the parameter a can be considered as the respondent's speed-accuracy trade-off. The speed-accuracy trade-off defines the respondent's performance at the IAT, and, once a trade-off is undertaken, it remains constant throughout the entire administration.

The starting point of information accumulation is expressed by parameter z. The location of the starting point affects the information accumulation process. If the starting point is closer to one of the two threshold, then less information is needed for giving the response associated to that threshold, generating a bias towards it. The parameter z measures this specific response bias.

The direction and the speed of the information accumulation process are expressed by

the the parameter v (i.e., drift rate). Drift rate determines both accuracy and speed performance of the respondents. The meaning of the drift rate can be considered both within and between respondents. When considered between respondents, it can be interpreted as the between–respondents differences in the decision-making processes. When considered within respondents and between experimental conditions, the drift rate expresses how the experimental condition affects the decision-making processes of each respondent. In the IAT case, the decision-making process is easier in the condition where the evaluative dimension and the target object with the strongest automatic association share the same response key than in the opposite condition.

Finally, the decision-making process is also influenced by preparatory operations which are not directly related to the decision itself, such as the preparatory movements for giving the response. The Parameter t_0 accounts for the nondecision component of the decision-making process.

In the IAT case, the automatic associations between target objects and evaluative dimensions positively affect the average information accumulation process (i.e., drift rate), resulting in faster and more accurate responses. This is true only for the condition consistent with the automatic associations. On the contrary, automatic associations negatively affect the drift rate, hence accuracy and speed performance, in the contrasting condition. The better performance usually observed in the compatible condition can be due to the use of the positive/negative intrinsic valence of the objects stimuli for their categorization. As such, the task-switching cost from attribute to target is avoided. Grounding on this observation, the DM allows for speculating that attitudes enter the IAT and influence respondents' performance through the object stimuli. This explanation is also in line with what found with the application of the ReAL model (Meissner & Rothermund, 2013).

The application of the DM to IAT data allows for separating processes that are directly and actively related to the decision process (e.g., drift rate, threshold separation) to those which are needed during the process but that mostly express preparatory operations (e.g., non decision component). Moreover, the information retrieved at the stimuli categories level is particularly useful to investigate whether and how attitudes and automatic associations are

affecting respondents' performance. Finally, the distinct response time distributions obtained from correct and incorrect responses can be further employed for gaining better insights on the processes underlying respondents' decisions and performance.

The application of the DM to IAT data comes with some drawbacks as well. A high number of incorrect responses is needed for estimating the response time distributions of error responses for each respondent. Otherwise, the estimates for the distributions of both correct and incorrect responses are not reliable. Consequently, respondents with a perfect performance have to be eliminated from the sample. However, as stated above, the IAT is an easy task, and a low percentage of error responses is usually observed. It follows that a large number of respondents would present data that are not suitable for a DM analysis. Another potential critical issue of the application of the DM to IAT data is that it is applied to each critical IAT block, namely each associative condition. For instance, in a Coke-Pepsi IAT two different DMs should be applied, one on the condition in which Coke (Pepsi) and Good (Bad) are associated, and one on the contrasting condition, where Pepsi (Coke) and Good (Bad) are associated. The application of separate DMs to each associative condition implies that the estimates obtained from the application to one critical condition cannot be directly compared with those obtained from the application of the DM on the opposite critical condition.

The task-switching cost can be inferred from the difference between the drift rate parameters in the two associative conditions. In the condition consistent with respondent's automatically activated association, the task-switching cost can be prevented by sorting the attitude objects according to their intrinsic positive or negative evaluation, in line with the recoding processes observed in Meissner and Rothermund (2013). Conversely, in the contrasting condition, the target objects cannot be sorted according to their intrinsic value anymore. Stimuli have to be sorted according to their nominal category, hence the task-switching cost from attribute to target object negatively affects the performance. However, the drift rate is not able to disentangle the process allowing for the prevention of the task switching cost from the automatic associations.

Finally, the DM is not able to yield the information provided by each individual stimulus but only to consider the information provided by the categories of stimuli.

3.2.2 The Discrimination-Association Model

The Discrimination-Association Model (DAM; Stefanutti et al., 2013) is a mathematical model based on the joint modeling of accuracy and time responses, specifically designed and developed for the IAT data.

The DAM assumes that each of the stimuli, irrespective of whether they are object stimuli or attribute stimuli, contains evidence for each of the four stimuli categories. This implies that the processing of each stimulus happens in parallel. The processes with which the evidence in favor of each category of stimuli is accumulated when a stimulus is presented can be conceived as independent Poisson processes. The independent Poisson processes are defined as *counters*, and they express the evidence in favor of a specific stimuli category contained by each process. Since the processing of the stimuli happens in parallel, the *counters* for all stimuli categories are activated when a stimulus is presented, and they compete between each other. The *counter* that wins the competition determines the observed response. Four *counters* (one for each category of stimuli) are hypothesized. For instance, suppose that a stimulus representing the *Coke* category in a Coke-Pepsi IAT is presented to a respondent. When the stimulus is presented, the *counters* for each of the stimuli category starts accumulating information for their respective category. If the *counter* for the *Good* category wins the competition, a correct response is observed in the Coke-Good/Pepsi-Bad condition, with its related response time. The same *counter* would end in an incorrect response in the contrasting condition (i.e., Pepsi-Good/Coke-Bad).

The DAM decomposes the IAT effect into three distinct processes: (i) stimuli discrimination (i.e., stimuli representativeness of their own category), (ii) automatic associations (i.e., associations between evaluative dimensions and target objects), and (iii) termination criteria (i.e., amount of information needed before a response is given). This model results in the estimation of three parameters, describing the processes into which the IAT effect is decomposed. The models parameters are the rates at which evidence is accumulated on each counter (i.e., stimuli discrimination and automatic association), and the termination criteria. The rates at which evidence is accumulated on each counter is expressed by the parameter λ_{ij}, where i

is the counter for each of the four categories of stimuli and j is the specific stimulus presented on the screen.

The stimuli discrimination parameter describes the strength of the association of the stimuli with their own category. Specifically, the discrimination rates describe the amount of evidence that target (resp. to attribute) categories accrue when target (resp. to attribute) stimuli are presented. As such, it expresses the ability of the stimuli to represent the category to which they belong. For instance, stimuli *Coke* of the Coke-Pepsi IAT are described by two values of λ, one describing the correct discrimination of the stimuli (i.e., $\lambda_{\text{Coke,Coke}}$) and one describing the incorrect discrimination of the stimuli (i.e., $\lambda_{\text{Other,Coke}}$). If the stimuli chosen for representing the target category *Coke* are prototypical exemplars of the category, the value of $\lambda_{\text{Coke,Coke}}$ is expected to be higher than the value of $\lambda_{\text{Other,Coke}}$.

The *automatic association* parameter directly derives from the association pattern between object stimuli and evaluative dimensions. The *automatic association* parameter regards the amount of evidence that target (resp. to attribute) categories accrue when attribute (resp. to target) are presented. The estimation of this parameter depends on the automatic association of each respondent. Taking the Coke-Pesi IAT as an example, if the respondent holds a preference for Coke, the automatic activation process facilitates the categorization task (i.e., higher accuracy and faster time responses) in the condition where *Coke* and *Good* share the same response key. Conversely, it impairs the categorization task in the condition where *Coke* and *Bad* share the same response key.

The *automatic association* parameter can help in disentangling the automatic association that drives the performance in each associative condition, and consequently, in clarifying the meaning of the IAT effect. Following the previous example, the *automatic association* parameter might highlight an high association between *Good* and *Coke*, along with a low association between *Good* and *Pepsi* and *Bad* and *Pepsi*. As such, it can be said that the performance is mostly driven by a positive evaluation of *Coke*, while *Pepsi* is associated with neither positive nor negative evaluations.

Termination criteria refers to the amount of evidence that needs to be accumulated before any response is given. The amount of information needed for producing the correct response

in the incompatible blocks (i.e., blocks against respondent's automatically activated associations) is usually larger. Consequently, the termination criteria are higher in the incompatible condition than in the compatible one. The termination criteria can hence be interpreted as either task difficulty or individual cautiousness.

The DAM provides useful information on the IAT functioning. Additionally, it overcomes some of the major issues of DM, namely the impossibility of obtaining reliable estimates when few or no errors are made and the separate application of the model to each critical block. Since the DAM assumes separate processes for correct and incorrect responses, few or no error responses only affect the estimation of the parameters concerning the processes leading to incorrect responses, while the parameters concerning the correct responses can still be reliably estimated. Moreover, the DAM is applied on the entire IAT data set, and termination criteria are the only parameter that varies across blocks, while in the DM both drift rates and threshold separation vary across blocks.

However, also the DAM presents some shortcomings. Stimuli discrimination provides important information on stimuli functioning and on their representativeness of the category to which they belong. This information is at the level of the stimuli category and not at that of the individual stimuli. Having information at the level of the individual stimuli would not only allow for testing their representativeness but also for delving deeper on the specific stimuli driving the IAT effect. Moreover, the way in which termination criteria have been conceptualized makes difficult to disentangle the respondent's contribution from that of the task in determining the observed response. This point is crucial for a better understanding of the IAT functioning. Specifically, there has to be a clear distinction between the properties of the task/stimuli, how they are affecting respondents' performance, and the respondents' characteristics mostly affected by the task characteristics.

3.3 Rasch Modeling

The models presented so far do provide interesting and useful information on the processes involved during the performance at the IAT. However, they overlook the information that

can be gathered from the individual stimuli used, which is their major pitfall. Indeed, as previous studies have pointed out (e.g., Bluemke & Friese, 2006), the characteristics of the stimuli (e.g., their representativeness of the category to which they belong) play a crucial role in the functioning of the IAT. As such, a modeling framework able to get a detailed information at the level of the individual stimulus would provide a fine-grained analysis of the IAT functioning based on its single, yet most important, components. By disentangling the contribution of the characteristics of the respondents from that of the characteristics of the stimuli, the Rasch model (Rasch, 1960) is able to provide such a fine-grained analysis at the individual stimulus level.

The Rasch model assumes that the variability at the level of the observed accuracy responses can be explained by a unique latent variable. Once the effect of this variable is accounted for, the correlation between the responses should be close to 0 (i.e., local independence). The observed response is the result of the interplay between respondents' characteristics, expressed by an ability parameter β, and item characteristics, expressed by a difficulty parameter δ. Therefore, the expected responses can be completely explained by just these two parameters that are the manifestations of the latent variable. A more thorough explanation of the Rasch model is provided in Section 4.1.

In the IAT case, the variability at the level of the observed responses cannot be completely exhausted by just respondents' ability and stimuli difficulty. Part of this variability can be ascribed to the associative conditions in which stimuli are presented.

The Many Facet Rasch Model (MFRM; Linacre, 1989) extends the Rasch model by allowing for other sources of variability (i.e., *facets*) to explain the variability at the levels of the observed responses. This approach already proved its usefulness for modeling IAT data across different domains of investigation (e.g., Anselmi et al., 2011; Anselmi, Vianello, Voci, & Robusto, 2013)

By using the MFRM, the associative conditions can be specified as a *facet* of the model. Therefore, the variability in the observed responses due to the associative condition is accounted for. Respondents, stimuli, and associative conditions are hence facets of the model, and they operate in concert for determining the likelihood of a response.

The MFRM is meant for estimation of the likelihood of categorical variables. As such, it cannot be directly applied to the response times of the IAT, which need to be discretized into ordered categories. Consequently, the model results in the estimation of the probability of giving the response within a time category. Usually, quantiles are used for the discretization of the continuous time responses. The number of quantiles into which the response times are divided is an *ad-hoc* choice made by the researcher.

Let k be a parameter describing the discretized scale of the response times, with $k \in \{0, 1, \ldots, m\}$. The MFRM for the analysis of IAT data takes on the form:

$$ln\left(\frac{P_{psck}}{P_{psc(k-1)}})\right) = \beta_p - \delta_s - \gamma_c - \tau_k, \tag{3.1}$$

where P_{psck} is the probability that respondent p (with ability β_p) would respond to stimulus s (with difficulty δ_s) in condition c at speed k. Parameter γ_c describes the easiness of condition c, while parameter τ_k describes the impediment of response k relative to $k-1$. Therefore, the additive effects of the speed of the respondent β_p, the speed of the categorization of the stimulus δ_s, the easiness of the condition γ_c, and the impediment of response k rather than $k-1$ define the probability that respondent p gives response k rather than $k-1$ to stimulus s in condition c.

By considering the IAT associative conditions as a *facet* of the model, it is possible to obtain either condition–specific stimuli estimates or condition–specific respondents' estimates. However, it is not possible to concurrently obtain condition–specific respondents and stimuli estimates because the model would not be identified. The estimates obtained in the former case allow for investigating whether the stimuli show a different functioning between associative conditions. By computing the difference between stimuli condition–specific estimates, it is possible to obtain a measure of the bias due to the associative conditions or, in other words, a measure of the contribution that each stimulus is giving to the IAT effect. The estimates obtained in the latter case allows for investigating the effect of the associative conditions on respondents' performances (if any).

Condition–specific stimuli estimates highlighted a positive primacy effect (e.g., Anselmi

et al., 2011, 2013), according to which the IAT effect is mostly driven by positive attributes (i.e., the *Good* exemplars are the stimuli that have the greatest difference in their estimates between the two associative conditions). This result made possible to interpret the IAT effect under a different perspective concerning both a Race IAT and a Weight IAT. In the Race IAT, the preference for White people observed on White respondents could be interpreted as the expression of ingroup preference rather than outgroup derogation. Similarly in the Weight IAT, Thin people tended to show an implicit preference for Thin people rather than Fat people. This result should be interpreted in light of the expression of ingroup preference rather than outgroup derogation.

In the IAT case, the information at the level of the individual stimulus provided by the MRFM can be used to investigate both its representativeness of the category to which it belongs and its specific contribution to the overall IAT effect.

At the respondents' level, the Rasch model allows for investigating whether respondents' performance (i.e., the indicator of the latent trait posited by the model) has been affected by the IAT associative conditions, and, if so, how and how much. Differently from the modeling approaches presented so far, Rasch modeling of IAT data provides important information at the individual stimulus level, which can in turn lead to a better understanding of the measure itself.

Unfortunately, this approach presents some drawbacks as well. Firstly, the MFRM applications presented in this section are all based on the discretized time responses. Besides a potential large loss of information, the discretization process presents an arbitrary component related to the decision on the number of quantiles to use. Results might change according to the number of quantiles in which the starting continue variable has been divided. Accuracy responses are not accounted for, hence the information that can be retrieved from them is overlooked. Moreover, since the focus was on the stimuli functioning between conditions, the difficulty of the two conditions was assumed to be the same across respondents. Consequently, it was not possible to investigate the bias due to the IAT associative conditions at the respondents' level. Finally, the fully-crossed structure was not accounted for, even though the remaining variability in the observed responses was acknowledged. However, it was entirely

ascribed to the IAT associative condition.

3.4 Common features, advantages, and drawbacks

Depending on the focus of the above mentioned models, they result in different and useful information regarding either the stimuli functioning (Rasch modeling) or the cognitive components involved in the performance at the IAT.

With the only exception of Rasch modeling, all modeling frameworks presented in this chapter point out that the performance at the IAT is not solely influenced by automatic processes but also by processes on which respondents have different levels of control. While other formal models are mostly focused on highlighting the processes underlying the performance at the IAT, Rasch modeling is more oriented on the information given by each stimulus, and how to use it for having a better understanding of the IAT effect.

The Quad model and the ReAL model do provide an interesting disentanglement of the IAT effect into the controlled and automatic processes governing the responses. They both point out the fact that the measure obtained from the IAT is not a process pure measure of implicit associations, but it also include a part of controlled processes which need to be taken into account for making meaningful inferences. However, they presented major shortcomings, starting from the fact that they use only a part of the information of the IAT, namely the accuracy responses. As already stated, the IAT is an easy task, and the observed error rates are usually not high enough to allow for a reliable estimation of model parameters. Consequently, some precautions have to be taken for performing analysis on IAT data under these frameworks. For instance, the ReAL model introduced a rtw that automatically adjust to the performance of each individual to increase the error rates. This expedient makes the parameters of the ReAL model obtained with this modified version of the IAT not directly comparable with the classic IAT scores obtained from traditional IATs. Moreover, to guarantee a high enough error rates for each combination of stimuli in associative conditions, the analysis for the Quad and the ReAL models are performed at the sample level or at the stimuli category level. However, the IAT was designed for investigating individual differences, hence

obtaining parameters that provide only a general information about the sample performance might not be in line with the original purpose of the measure itself.

Both the Quad model and the ReAL model posit parameters describing the stimuli (i.e., parameters D, L and stimuli discriin the Quad model, the ReAL model, respectively) and parameters for describing the automatic associations (i.e., parameters AC and A in the Quad model and the ReAL model, respectively). However, the parameters of the Quad model and those of the ReAL model describes the two components at the level of the entire IAT. The DAM overcomes this issue by providing a more detailed analysis at the level of the categories of stimuli.

The ReAL model, the DM, and the DAM highlighted how the categorization task can be simplified by exploiting the positive and negative valence triggered by a target stimulus. Both the DAM and the ReAL model postulate the activation of another stimulus category when the stimulus belonging to a category is presented. Among other differences, the DAM and the ReAL models also differentiate themselves according to the direction they assume for the activation of the recoding strategy. The activation of the evaluative dimension when a target stimulus is presented, and not the other way around, is the working assumption on which the ReAL model is based. Consequently, only target objects can be recoded and categorized according to their positive/negative value, while the same cannot be done for the evaluative attributes. The DAM goes beyond because it does not make such a strong assumption on the activation of the automatic associations of the stimuli. According to the DAM, each stimulus contains information regarding the other categories. As such, each target object can accrue information regarding both the evaluative dimensions and the opposite target object. This holds true also for the stimuli representing the evaluative dimensions. Each evaluative attribute can accrue information regarding the categories of both target objects and the opposite evaluative dimension. This makes the DAM a more flexible model than the ReAL model. Moreover, it allows for empirically testing the basic assumption on which the ReAL is based.

The DM does not explicitly mention a recoding process, but it refers to a task-switching cost. The difference in the drift rates between associative conditions indicates that the cate-

gorization task is easier in one condition than the other. Klauer et al. (2007) speculate that the facilitation effect of the associative condition can be attributed to the categorization of the target stimuli according to their positive/negative valence. By doing so, the task-switching cost from attribute to target is prevented, ending up in a better performance. This strategy facilitates the categorization task in the condition consistent with one's own automatic association, while it hinders it in the condition against the automatic association. The facilitation (hindering) effect of the task-switching cost can be seen in the drift-rates difference between the conditions. Similarly to the ReAL model, the categorization task is simplified only by exploiting the intrinsic values of the target objects and not the other way around. In the DM, this is consistent with the assumption of the serial processing of the stimuli. Additionally, Klauer et al. (2007) clearly state that attitudes influence the performance at the IAT through the target objects. An issue of the DM is that it cannot disentangle what part of the performance is driven by this recoding, controlled process, and which is actually ascribable to the activation of automatic associations.

Another potential critical aspect could be the complexity of these models, specifically of the DM and the DAM. While both these models do provide a more complete information on the IAT effect, hence a better understanding of the measure itself, their understanding is not straightforward. To gain a deep understanding of the models themselves, and on the clear advantages related to their use, users are required to have at least a basic knowledge on random walk processes (DM) or Poisson processes (DAM). Unfortunately, this kind of expertise is not widespread among researchers using the IAT. This might prevent them from using these models and discard them in favor of a simpler, but less sound, approach.

The shortcoming of these models have already been highlighted and discussed. One common drawback of all of them is that they cannot provide any information at the stimuli level. However, the Rasch model applications stressed the importance and usefulness of having such an information not just for the investigation of the stimuli functioning itself, but also for a better understanding of the IAT measure. On the other hand, the applications of the Rasch model to IAT data that have been attempted so far are based on only the (discretized) time responses. As such, the information from both accuracy responses and the continuous

nature of time responses is lost.

The common and most outstanding drawback of these models is that none of them is accounting for the fully-crossed structure of the IAT data described in Section 1.4. Consequently, the sources of random variability in the data are left free to bias the estimation of the parameters.

Chapter 4

Rasch model, log-normal model, and their specification for analyzing IAT data

This chapter is organized in two main sections. In the first section, the Rasch model and the log-normal model are briefly outlined. Then, the similarities between the Rasch model and the Generalized Linear (Mixed-Effects) model are described. The procedure for the estimation of the Rasch model parameters by using Generalized Linear Mixed-Effects Models (GLMMs) with a *logit* link function is illustrated, as well as the procedure for estimating the log-normal model parameters from Linear Mixed-Effects Models (LMMS).

In the second section, the random structures of the GLMMs and the LMMs used for estimating the Rasch model and the log-normal parameters from IAT accuracy and log-time data, respectively, are presented. Three random structures for accuracy responses (Rasch model), as well as three random structures for log-time responses (log-normal model) are introduced. The first one is the simplest one, and it is considered as the Null model against which the models with the other two random structures are compared. The second and third models have the same level of complexity. They differentiate each other according to the random factor on which the multidimensionality of the associative condition is allowed, that can be either the respondents (Model 2) or the stimuli (Model 3). The best fitting model, and consequently the Rasch model and the log-normal model parameters that can be estimated,

depend on the observed data.

For illustration purposes, the Rasch model is initially presented with the typical notation for its parameters, namely β, indicating persons' abilities, and δ, indicating item difficulties. However, since also the item parameter of the log-normal model is indicated with δ, a different notation for the Rasch model is employed. The new notation of the Rasch model resembles the one typically used in Item Response Theory in general. The respondents' parameters are indicated with the Greek letter θ. The item parameters are denoted with the Latin letter b. The respondents are indicated with the subscript p ($p \in \{1, \ldots, P\}$) and the stimuli/items with the subscript s ($s \in \{1, \ldots, S\}$). In the specification of Linear Mixed-Effects Models, the single observation on each respondent p on each stimulus s in each associative condition c ($c \in \{1, \ldots, C\}$) is indicated as i ($i \in \{1, \ldots, I\}$).

4.1 Modeling dichotomous responses

According to Item Response Theory (IRT) models, the observed response to an item can be explained by a common characteristic shared by both the person and the item, which lie on the same latent trait (DeMars, 2010). IRT scoring accounts for the moderation of item characteristics in explaining the relationship between the person's latent trait, often identified with θ, and the observed response. IRT models can be distinguished according to the number of parameters used for describing the item characteristics (e.g., DeMars, 2010).

The simplest model is the 1-Parameter Logistic model (1PL, Equation 4.1). The 1PL model ad the Rasch model (Rasch, 1960) are mathematically equivalent. According to the 1PL model, the probability of a correct response to an item is a function of the respondent's characteristic θ and an item characteristic, defined as difficulty, b:

$$P(x_{ps} = 1|\theta_p, b_s) = \frac{exp(\theta_p - b_s)}{1 + exp(\theta_p - b_s)} \tag{4.1}$$

The difficulty b is defined as the amount of latent trait θ that a person needs for having a higher probability of choosing the correct response over the incorrect response.

The 2PL model (Equation 4.2) (Birnbaum, 1968) also considers the influence of each item discrimination power (parameter a) in explaining the relationship between the respondent's ability and the observed response:

$$P(x_{ps} = 1|\theta_p, b_s, a_s) = \frac{exp[a_s(\theta_p - b_s)]}{1 + exp[a_s(\theta_p - b_s)]} \tag{4.2}$$

As it can be seen from Equation 4.2, parameter a changes the relationship between respondent's parameter θ and the item difficulty parameter b. The larger the value of a_s, the lower the overlap between the distributions of the response variables of two respondents with different values of θ. In this sense, parameter a_s can be interpreted as the discriminating power of the item. Items with large value of a_s are best able to discriminate between respondents with different levels of θ.

Both the 1PL and the 2PL models assume a lower asymptote at 0 (i.e., the value taken by the function as θ approaches $-\infty$) and an upper asymptote of 1 (i.e., the value taken by the function as θ approaches $+\infty$). The Rasch model assumes a lower asymptote at 0 as well. The assumption of the lower asymptote approaching zero implies that respondents with extremely low levels of ability have an extremely low probability of endorsing the correct response. Conversely, assuming an upper asymptote of 1 implies that respondents with extremely high levels of ability have an extremely high probability of endorsing the correct responses.

However, there might be cases in which respondents with an extremely low level of ability endorse the right response just out of luck (lucky guess), or that respondents with an extremely high level of ability endorse the incorrect response just out of distraction (careless error). In the first case, the lower asymptote cannot approach zero anymore, since even respondents with a level of ability that approaches $-\infty$ have a probability of providing the correct response higher than 0. In the latter one, the upper asymptote has to be moved downward because even respondents with a level of ability that approaches $+\infty$ have a probability of correctly endorsing the correct response lower than 1.

The 3PL and 4PL models have been introduced for modeling these occurrences, respectively.

The 3PL model (Equation 4.3) (Lord, 1980) adds a third parameter (c) to explain the response behavior:

$$P(x_{ps} = 1|\theta_p, b_s, a_s, c_s) = c_s + (1 - c_s)\frac{exp[a_s(\theta_p - b_s)]}{1 + exp[a_s(\theta_p - b_s)]}, \tag{4.3}$$

where c_s is the probability that a respondent with a low level of ability guesses the correct response. Parameter c_s hence moves upward the lower asymptote, and it represents the probability that a respondent with an extremely low ability will correctly answer an item with difficulty b

The 4PL model (Equation 4.4) (Barton & Lord, 1981) adds a fourth item parameter (e) to describe the response behavior:

$$P(x_{ps} = 1|\theta_p, b_s, a_s, c_s, e_s) = c_s + (e_s - c_s)\frac{exp[a_s(\theta_p - b_s)]}{1 + exp[a_s(\theta_p - b_s)]}, \tag{4.4}$$

where e_s represents the probability that a respondent with an extremely high level of ability will incorrectly answer an easy item (i.e., careless error). As it can be seen from $(e_s - c_s)$ of Equation 4.4, the upper asympote is defined by parameter e_s.

4.1.1 The Rasch model

Despite the 1PL IRT model and the Rasch model are mathematically equivalent, the notational system used for their parameters is different. In the Rasch model, the item parameter is described by the Greek letter δ and the person's parameter is described by β. In this section, the typical notation of the Rasch model is used. However, in Section 4.4 the notation typical of IRT models is used to distinguish the Rasch model parameter estimates from the estimates of the log-normal model and those of the GLMMs.

The starting point for the development of the dichotomous Rasch model (Rasch, 1960) involves the engagement of a person p on an item s to produce a response x_{ps} (Andrich & Marais, 2019). The engagement between the person and the item results from a single variable that is a common property shared by both the person and the item. The item variable

is supposed to trigger the same person variable in all respondents. For instance, for assessing mathematics proficiency the items must contain some degree of mathematics ability. To give the correct response, persons must engage with the mathematics proficiency required by the item.

The engagement between persons and items results in the observed responses x_{ps}, which can be represented in a P ($p \in \{1, \ldots, P\}$, persons) $\times$ items S ($s \in \{1, \ldots, S\}$, items) response matrix $\boldsymbol{X}$ (Table 4.1).

Table 4.1: Response matrix $P \times S$, starting point for estimating the Rasch model.

	Items						
	1	2	...	k	...	s	
1	x_{11}	x_{12}	...	x_{1k}	...	x_{1k}	r_1
2	x_{21}	x_{22}	...	x_{2k}	...	x_{2k}	r_2
$\vdots$	...	...	...	...	...	...	$\vdots$
Persons v	x_{v1}	x_{v2}	...	x_{vk}	...	x_{vk}	r_v
$\vdots$	...	...	...	...	...	...	$\vdots$
p	x_{p1}	x_{p2}	...	x_{pk}	...	x_{ps}	r_p
	s_1	s_2	...	s_k	...	s_s	

Each cell represents the response of person p to item s. The response is a dichotomous response that can take only the values $x_{ps} = 0$ (incorrect response) or $x_{ps} = 1$ (correct response).

The across-columns sum r_p (i.e., number-correct) represents the total score of each respondent (i.e., the total number of correct responses given by the respondent), regardless of the specific pattern with which the correct responses were given. The number-correct is a sufficient statistic for estimating the person's parameter β (Wright & Stone, 1979; Wright, 1997). Two respondents might have the same number-correct obtained with different patterns of correct responses. Since the specific pattern does not matter for the determination of the number-correct, the two respondents with same number-correct obtained with different pattern will have the same person estimate β. This feature is one of the peculiarities that distinguish the Rasch model from other IRT models. For instance, in the 2PL two respondents with the same number-correct might not have the same level of θ because the relationship

between the item and the respondent's estimate is moderated by the discrimination parameter a.

Similarly, the across-rows sum s_s (i.e., proportion-correct) represents the total score of each item (i.e., number of correct responses obtained by each stimulus), regardless of the specific pattern with which the responses were obtained. The proportion-correct is a sufficient statistic for estimating the item difficulty parameter δ. Two items might have the same proportion-correct resulting from different pattern of responses, but since the specific pattern is not relevant for the determination of the proportion correct, the two items will have the same item estimate δ. The concept of sufficient statistics is directly related to that of Local independence, which is further illustrated in paragraph $4.1.1$.

The observed response in each cell x_{ps} hence depends on both persons' characteristics and items characteristics. Characteristics of both persons and stimuli can be located on a specific point of the latent trait, which is the common variable they share. The locations of each respondent p on the latent trait are described by parameter β_p. The location of each stimulus on the latent trait is described by parameter δ_s. While the observed response for each combination of $p \times s$ can take only the value 0 and 1, the parameters β_p and δ_s can take any real values from $-\infty$ to $+\infty$.

The fact that persons and items are located on the same latent traits is the great advantage of the Rasch model. By sharing the same latent trait, it is possible to directly compare persons' estimates with items estimates, and hence a measure of the distance between them can be obtained. Therefore, it is possible to predict the probability that a person with a certain level of β has of correctly respond to an item with a certain level of δ. Since the observed response is a function of respondents' and stimuli characteristics located on the same latent trait, it is possible to speculate that a respondent would correctly respond to stimuli below his/her level of ability β_p (i.e., the probability of a correct response is higher than 50%):

$$\text{If } (\beta_p - \delta_s) > 0 \text{ then } P(x_{ps} = 1) > 0.50. \tag{4.5}$$

Also the opposite holds true. When the location of the item is above the location of the

person, the probability that a correct response is given is below 50%:

$$\text{If } (\beta_p - \delta_s) < 0 \text{ then } P(x_{ps} = 1) < 0.50. \tag{4.6}$$

Assuming that respondents and items lie on the same latent variable implies that the probability of a correct response can be related with the difference between respondents parameters and item parameters, as shown in Equation 4.5 and Equation 4.6.

However, the probability of a correct response is bounded between 0 and 1, while the parameters, and hence their difference, can vary between $-\infty$ and $+\infty$. The difference between respondents' and items parameters can be forced to only positive numbers by using the exponential distribution:

$$0 \leq exp(\beta_p - \delta_s) < +\infty \tag{4.7}$$

Still, the difference between person's ability and item difficulty cannot be used to predict a probability, because it is allowed to take any positive value from 0 to $+\infty$. To map the difference between respondents and item parameters on the same scale of the probability, and hence to use them for predicting the probability of a response given respondents' ability and item difficulty, Equation 4.7 can be standardized by $1 + exp(\beta_p - \delta_s)$. The probability for a correct response for a given β_p and a given δ_s can hence be expressed as:

$$P(x_{ps} = 1|\beta_p, \delta_s) = \frac{exp(\beta_p - \delta_s)}{1 + exp(\beta_p - \delta_s)}, \tag{4.8}$$

which is the typical formulation of the Rasch model for the probability of a correct response.

As said before, the Rasch model was originally formulated in terms of odds and *log-odds*. Equation 4.8 can hence be rewritten in terms of *log-odds*:

$$\beta_p - \delta_s = \ln\left(\frac{P(x = 1|\beta_p, \delta_s)}{1 - P(x = 1|\beta_p, \delta_s)}\right), \tag{4.9}$$

or, by applying the properties of logarithms to Equation 4.8, the Rasch model can be rewritten

as:

$$P(x_{ps} = 1|\beta_p, \delta_s) = \frac{1}{exp(\delta_s - \theta_p)} \tag{4.10}$$

The formulation in Equation 4.8 makes clear that the only thing that matters for the estimation of the expected probabilities is the difference between β_p and δ_s. This difference expresses the distance between the location of respondent p from the location of stimulus s on the latent trait. The probability of a correct (incorrect) response changes according to the distance between respondent's and item locations. The probability of a correct response is 50% when the respondent's location is equal to the item location. The variance for the the expected probabilities of responses when the locations of the respondent and that of the stimulus corresponds is maximized.

The more the location of a respondent on the latent trait is above the location of the item, the higher the probability of observing a correct response (see Equation 4.5). Similarly, the less the distance between the location of the respondent and that of the item on the latent trait, the lower the probability of observing a correct response. Also the opposite holds true. The more the location of the respondent is below the item location, the higher the probability of an incorrect response (see Equation 4.6), and, conversely, the less the location of the respondent is below the location of the item, the lower the probability of an incorrect response. The relationship between the respondent's parameter and the item parameter defines the cumulative nature of the Rasch model.

The Rasch model assumes a logistic probability function, and the measurement units of the respondents' parameters, the item parameters, and their difference, are the *logits* (i.e., *log-odds* of the probability of giving the correct response).

The Rasch model is based on three main assumptions, namely linearity, comparison invariance, and local independence. These assumptions are briefly outlined in the following paragraphs, with a specific focus on conditional independence and on the consequences of its violation.

Linearity of the scores. The linearity of the scores is obtained with the logarithm transformation of the odds. By applying this transformation, person's parameters and item parameters are placed on the same continuous latent latent trait. The measurement units of the latent trait are the *logits*, which define an interval scale for the interpretation of the scores.

This linear transformation allows for setting the lowest parameter observed equal to 0, without losing the original relationship between the estimates. Consequently, comparison invariance (described in the following paragraph) assumption is satisfied.

Comparison invariance. The comparison between any two persons is independent from the set of stimuli on which the comparison is based, as well as independent from the comparison between any other two persons. The same holds for the stimuli, so that the comparison between any two items is independent from the respondents on which the comparison in made, as well as from the comparison between any other two persons.

The comparisons are invariant in the sense that the comparison between persons only depends on the ability parameters of those two persons, and the comparison between stimuli only depends on the stimuli properties.

Local independence. According to the Rasch model, a person with a level of ability β greater than the item difficulty δ has a greater probability of responding correctly than incorrectly to that item. Conversely, a person with a level of ability β lower than the item difficulty δ has a higher probability of responding incorrectly than correctly to theta item. The variability between the item responses can hence be explained in terms of the person's level of ability β. As such, ability β can be considered as the source of general dependence between the items, and, once it is accounted for, any other relationship between the items should disappear. The capacity of the persons' parameters β to explain all the variability between the responses, is called local independence (Andrich & Marais, 2019). In other words, local independence assumes that, once the effect of person parameter is accounted for, any other relationship between the stimuli should disappear.

The statistical independence of responses implies that the probability of correctly re-

sponding to different items is equal to the product of the probabilities of answering each of them correctly. The local independence of the responses can be formalized as follows:

$$P(\boldsymbol{X}) = \prod_{p} \prod_{s} P(x_{ps}), \tag{4.11}$$

where $\boldsymbol{X}$ is the $P \times S$ matrix of the responses.

The violation of local independence can happen in two main instances, either by involving multidimensionality or response dependence. The consequences of the local independence violation due to either multidimensionality or response dependence move in opposite directions but they both result in less reliable parameters estimates and predictions.

Unidimensionality posits that item responses are explained by only one latent trait dimension, shared by both respondents and stimuli, and it is the basic underlying assumption of the Rasch model. Multidimensionality refers to those cases where there are unexpected person's parameters other than β involved in the responses to the items.

Multidimensionality is indeed a property of many different scales used for psychological assessment. For instance, the Big Five Questionnaire (Caprara, Barbaranelli, Borgogni, & Perugini, 1993) is a questionnaire for the assessment of the Big Five personality traits, composed of different subscales. The items in each of the 5 subscales are aimed at assessing one of the 5 personality traits posited by the Big Five theory (i.e., agreeableness, extraversion, openness to experience, neuroticism, conscientiousness), and they can be grouped according to the personality trait they aim for. As such, they show a between–subscale variability which cannot be explained by only the person's parameter β.

Multidimensionality can also raise from stimuli linked by common attributes such as a common item stems, common stimulus materials, or common item structures (Andrich & Marais, 2019). Consequently, stimuli will display a variability that cannot be understood just in terms of ability parameters β.

Response dependence (i.e., for a fixed person, hence for a fixed level of ability β, the response to an item might depend to the response to a previous item; Andrich & Marais, 2019) violates the local independence assumption as formulated in Equation 4.11. The probability

associated to the responses to each item are not independent events anymore, hence their probabilities cannot be multiplied. Consequently, it is not possible to state that the probability of correctly responding the entire set of items corresponds to the product of the probabilities of responding correctly to each of them.

Response dependence can happen when the response given to an item is used as a clue for responding to the following item. Response dependence might also arise during the performance at computerized task, when the response to a previous stimulus might leave a carry-over effect on the response to the following stimulus (Westfall et al., 2014). Specifically, in the second case, the variability at the item level is affected by new sources of variability which are mostly composed of error variance.

Violating local independence affects the fit of the data to the model (Andrich & Marais, 2019), and produces unreliable parameter estimates (e.g., Barr et al., 2013; Judd et al., 2012). When local dependence is due to multidimensionality, extra sources of random noise are added to data. The error variance is hence increased, producing less accurate and reliable predictions. When local dependence is due to response dependence, the similarity of responses of persons across items is higher, leading to a lower error variability in the data.

4.2 Modeling time responses

By modeling response times within an IRT approach, an interaction between the parameters defining persons' accuracy responses and time responses is implicitly assumed. This is nothing else than the speed-accuracy trade-off also reported in previous analysis of the IAT data (e.g., Klauer et al., 2007).

Traditionally, in IRT modeling the speed-accuracy trade-off has been expressed by adopting a regression parameter for respondents' ability on their response times. Consistently with IRT models in general, also items are fundamental in determining the time responses. It is commonly assumed that more difficult items do need for more time to get a response. An item parameter is needed to describe the time absorbing power of the item (van der Linden, 2006).

4.2.1 The log-normal model

A log-normal model for the analysis of the response times to a test has been introduced by van der Linden (2006). This model is part of a hierarchical model for the modeling of accuracy and time responses in an IRT framework (van der Linden, 2009). As also stated by van der Linden (2006) and van der Linden (2009), the models used for accuracy responses (i.e., any of the IRT models presented in the previous section) and for time responses can be employed separately for the analysis of accuracy and time responses to a test. The advantage of using the hierarchical approach in van der Linden (2009) is that the relationship between the parameters of the IRT model and and the time parameters can be studied and understood at a second (combined) level of modeling.

The log-normal model, as its name suggests, assumes a normal density distribution for the logarithm of the time responses. The use of a log-normal family can be traced to its good fit to the observed data already observed in previous work (e.g., Thissen, 1983; van der Linden, 2006). More trivially, it comes natural to model with a normal distribution (defined over the entire real continuum) the log transformation of a variable that is a non-negative variable by definition (the response times) (van der Linden, 2006).

The structure of the original formulation of the log-normal model is analogous to the one of the 2PL IRT model in Equation 4.2 for mainly three reasons.

First, both the 2PL and the log-normal models impose the same structure on the mean of the distribution of the binary response variable and on the mean of the distribution of the continuous variable, respectively. In both cases, the mean is represented by the difference between the respondents' and the items parameters operating in opposite directions.

Second, both models assume a parameter that changes the relationship between the item parameter and the respondent's parameter, namely a discrimination parameter. Further details on the effect and the interpretation of the discrimination parameter are illustrated after the mathematical specification of the log-normal model.

Finally, given the nature of the distribution of the response times (it is bounded at 0), the log-normal model does not need the definition of a lower asymptote (i.e., a guessing

parameter like in 3PL model in Equation 4.3).

The response time t (i.e., the realization of a random variable T) of a person p to an item s can be expressed by positing the normal density distribution of the log-response time:

$$f(t_{ps}|\tau_p, \delta_s) = \frac{\alpha_s}{t_{ps}\sqrt{2\pi}} exp\left\{-\frac{1}{2}[\alpha_s(ln\ t_{ps} - (\delta_s - \tau_p))]^2\right\}. \tag{4.12}$$

As in IRT models, both respondents' parameters τ_p and δ_s are allowed to vary between $-\infty$ and $+\infty$. Although the sign of the persons' parameters is reversed, the mean of the distribution of Equation 4.12 resembles the one of the 2PL in Equation 4.2. The change in the sign of respondents' parameters allows for interpreting the parameter as a speed parameter, according to which, the larger the value of τ_p, the faster the responses given across items (i.e., the respondent tends to spend less time on the items).

Parameter δ_s describes the time intensity (or time consumingness) of an item, which is the time the stimulus requires to be responded. The larger the value of δ_s, the higher the amount of time respondents need to give the response. Parameter α (i.e., the reciprocal of the standard deviation of the normal distribution) is the discrimination parameter of the model. A larger value of α_s means less dispersion for the log-response time distribution on item s. Consequently, it can be said that the item has a better discrimintating ability between different respondents with different levels of speed. Parameter α_s affects the relationship between respondents' speed τ_p and item time intensity δ_s, similarly to what happens when parameter a_s changes in the 2PL model. If the value of α_s increases, the distributions of the log-time for any two values of the speed parameter show less overlap.

The 1PL model in Equation 4.1 can be considered as a constrained model deriving from the 2PL model in Equation 4.2. The constraint is imposed on the discrimination parameter a, which is forced to be equal across all items (1PL, Rasch model). A similar reasoning can be done for the log-normal model, by forcing α_s to be one for all items s ($\alpha_s = 1$ for all s, $s \in \{1, \dots S\}$). These constraints bring a parametrization similar to that of the 1PL/Rasch model. In the empirical application in van der Linden (2006), both a non-constrained and a constrained version of the log-normal model were tested in terms of goodness of fit to the

data. The normal analogs of the non-constrained and constrained models were tested as well. The normal models were the ones showing the worst goodness of fit, while constraining α to be equal across all items did not affect much the goodness of fit of the log-normal models.

4.3 Linear Mixed Effects Models

As illustrated in the previous sections, an IRT (or Rasch) approach for modeling both accuracy and time responses provides a detailed information on the parameters that determine the observed responses.

The use of a log-normal model can indeed overcome the issue related to the discretization of the response times for the application of the Many Facet Rasch Model in Section 3.3. The use of a separate model for accuracy responses, always under an IRT or Rasch framework, allows for obtaining useful and detailed information also from accuracy data, with a similar parameterization as that obtained from the log-time responses. Potentially, the estimates obtained from the two models can be combined at a second level of modeling by using a hierarchical approach as that illustrated in van der Linden (2009).

Despite this approach sounds promising, it cannot account for the fully-crossed design of IAT data, and the sources of dependency related to it. As thoroughly illustrated in the introduction, the fully-crossed design of the IAT comes with several sources of variability at different levels. These sources of variability generate dependencies at the level of the single observations, which violate the assumption of conditional independence. Conditional independence is not only a necessary assumption for the application of the Rasch model, but a basic assumption needed for obtaining reliable results with any statistical analysis. Violating the assumption of conditional independence brings to biased parameter estimates which can in turn lead to an inflated probability of committing Type I error or to an underestimation of the importance of the experimental condition (Barr et al., 2013; Judd et al., 2012; McCullagh & Nelder, 1989).

Linear Mixed Effects models (LMMs) are the most straightforward way to deal with this data structure. Moreover, both respondents and stimuli can be conceptualized as random

factors by specifying the appropriate random structure. As such, the issues concerning the decision to follow either a *by-participant* approach or a *by-stimulus* approach for performing the analyses is overcome.

LMMs allow for decomposing the error variance by specifying an appropriate random structure with different levels, that are the levels where uncontrolled random variation can be reasonably found (Doran, Bates, Bliese, & Dowling, 2007). The error variance can be partitioned into random effects that reflect the assumption on the structure of dependency created by the random variability at different levels. By doing so, the multilevel structure of the data, which reflects the random variability of the population from which the various levels are drawn, is accounted for (Barr et al., 2013).

Finally, LMMs can be applied to both continuous data, such as the log-transformation of the response times, and to dichotomous responses. In the latter case, a Generalized Linear Mixed-Effects Model (GLMMs) is needed, with the appropriate link function expressing the relationship between the linear combination of the predictors (i.e., linear component of the model) and the observed response.

4.3.1 Generalized Linear Mixed-Effects Model and Rasch Model

In a Generalized Linear Model (GLM), the linear predictors are not directly related with the observed response. They need to be linked together with a specific function, which goes under the name of *link function*. The type of link function that needs to be used depends on the nature of the observed variables (McCullagh & Nelder, 1989).

For the illustration of the structure of the GLM, and of its expansion for including random effects, we focus on the case of binomial responses $x_{ps} \in \{0, 1\}$, describing the accuracy responses at the IAT.

The linear combination of the predictors is defined by the form of the model expressed by the model matrix $\boldsymbol{X}$, and it determines the linear component of the model. The linear component is defined for each cell of the $P \times S$ $\boldsymbol{X}$ matrix, and it is denoted with η_{ps}.

The natural link function g that relates observed binomial responses with the linear com-

ponent of the model η_{ps} is the *logit* (the logarithm of the odds, McCullagh & Nelder, 1989), and it yields a probability value μ_{ps}:

$$\eta_{ps} = logit(\mu_{ps}) = ln\left(\frac{\mu_{ps}}{1-\mu_{ps}}\right), \tag{4.13}$$

where μ_{ps} is the probability of a correct response associated to each observed response x_{ps}.

Each link function is an invertible function, and the inverse for the *logit* link function is expressed as:

$$\mu_{ps} = logit^{-1}(\eta_{ps}) = \frac{1}{1+exp(-\eta_{ps})}. \tag{4.14}$$

The structure of the inverse *logit* link in Equation 4.14 can be equated to the Rasch formulation in Equation 4.10, and, consequently, it is possible to obtain a Rasch parametrization of the data by using a GLM on binomial responses with a *logit* link function (De Boeck et al., 2011; Doran et al., 2007; Gelman & Hill, 2007).

From now on, to distinguish the parameters of the Rasch model from the parameters obtained with the LMMs the former ones will be denoted with θ_p and b_s, and they will refer to respondents' ability and item difficulty, respectively.

When there are reasons to believe that sources of variability can generate dependencies between the observations, such as in the IAT case, random effects accounting for the random factors generating the uncontrolled random variability should be included in the model matrix of the linear component. By doing so, the error variance is partitioned in different levels that are defined by the factors considered as random. The partitioning of error variance into specific factors makes it controllable and accountable for (Doran et al., 2007).

The $\boldsymbol{X}$ matrix that defines the linear component of the GLM needs to be extended to include the random factors. The linear component hence takes on the form:

$$\eta = \boldsymbol{X}\beta + \boldsymbol{Z}d, \tag{4.15}$$

where β indicates the coefficients for the fixed effects, $\boldsymbol{X}$ is the model matrix of the fixed effects β, $\boldsymbol{Z}$ is the $P \times Q$ matrix of the random effects (i.e., Q is the dimension of the random

effects vector), and d is the vector of random effects predictors.

The dimension q of d is defined by the number of levels of each random factor, and their eventual combinations. For instance, if respondents are specified as a random factor, the dimension q of the random effects vector will have as many levels as the number of respondents. Consequently, the dimension of d can be potentially very large (Doran et al., 2007). The distribution of the random effects is estimated as a multivariate normal distribution (i.e., $\mathcal{MVN}$) with mean 0 and a $Q \times Q$ variance-covariance matrix $\mathbf{\Sigma}$, which is determined by a single vector parameter Γ (Doran et al., 2007). The dimension of Γ is usually rather small, and its size is determined by the number of random factors specified in the model, regardless of the number of levels they include. For instance, consider a model in which respondents' variability, items variability, and items variability in three different conditions are accounted for. In this model, five random factors are specified, one for respondents' variability, one for items variability, and one allowing for the multidimensionality of the stimuli variability in the three conditions. The dimension of the vector parameter Γ for this model is 5, and it remains 5, regardless of the number of respondents or items used.

The objective of LMMs is then to estimate the parameters of the fixed effects as defined by vector β and the parameters of the random effects, defined by vector Γ. Consequently, the parameters estimated for the random factors are not the parameters associated to each level of each factor, but the variance of the populations from which the random factors are drawn. This is the reason why d is indicated with a Latin letter, because it does not indicate population parameters.

Nonetheless, a measure for each level of each random factor is obtained in the form of *conditional modes*, which are the values that maximize the conditional density of the random effects given the vector of parameters (fixed and random) and the observed data (Doran et al., 2007). The conditional modes that describe the deviation from the fixed factors of each level of each  om factors are meta-parameters (Pastore, 2015), and are usually referred to as *Best Lineaer Unbiased Predictors* (BLUP, Pinheiro & Bates, 2006).

BLUP are used for the estimation of the Rasch model parameters.

Concerning the stimuli, the easiness estimates b_s are obtained by adding the conditional

mode of each stimulus, considered as a random factor, to the estimates of the fixed effects. In the IAT case, the higher the value of stimuli easiness b_s, the easier the stimulus, meaning that it is easily recognized and sorted to the category to which it belongs.

Similarly, adding the conditional mode of each respondent to the estimates of the fixed effects results in the respondents' ability estimates θ_p. In the IAT case, the higher the value of θ_p, the higher the ability of the respondent in correctly categorizing the stimuli.

In the Rasch model, the respondents' parameters and the stimuli parameters move in opposite directions. When the respondents' parameters and the stimuli parameters are obtained by using the GLMMs, their estimates move in the same direction, hence resulting in an additive effect. The item parameter b_s can no longer be interpreted as an impediment property (difficulty) of the item but it should be interpreted as a facilitation property of the stimulus (easiness) (De Boeck et al., 2011; Doran et al., 2007). When both θ_p and b_s are high, then the probability of a correct response is high. When high values of θ_p are combined with low values of b_s, the probability of a correct response for each respondent is as much penalized as their ability cannot balance out easiness of the stimulus.

Log-normal model estimates. The Rasch model estimates can be obtained by combining together the fixed component and the random component of the GLMMs applied on accuracy responses. Instead of being governed by the difference between the respondents' ability and the stimuli easiness, the probability of a correct response is governed by the additive effect of the respondents' ability and the item easiness. Consequently, the interpretation of the stimuli parameters b_s changes.

In a similar vein, log-normal model estimates can be obtained by combining the fixed factors to the random factors of the LMMs applied to the log-time responses. In the typical formulation of the log-normal model (Equation 4.12), the mean of the distribution of the expected log-time responses is expressed by the difference between the time intensity parameters of the stimuli δ_s and the speed parameters of the respondents τ_p (i.e., $\delta_s - \tau_p$). In the LMMs, the mean of the distribution is defined by the additive effect between the respondents' and the stimuli characteristics, which move in the same direction. Consistently, the lower the

value of speed parameter τ_p, the higher the speed, and the lower the value of δ_p, the lower the time each stimulus requires for getting a response.

When respondents with a low value of τ_p (i.e., high speed) respond to items with a low δ_s (i.e., low time intensity), the response times are fast. When a respondent with a low value of τ_p encounters a stimulus with a high value of δ_s, th speed of the response depends on the distance between the respondent's speed and the item time intensity.

4.4 Random structures

The random structures of the GLMMs and those of the LMMs are the same. The features differentiating the models are the assumptions on the error term ε and the dependent variable. In the GLMMs, the error term is supposed to follow a logistic distribution (i.e., $\varepsilon \sim \mathcal{L}(0, \sigma^2)$, where $\mathcal{L}$ is used to denote the logistic distribution of the disturbance as in Doran et al. (2007)) and the dependent variable is the accuracy response to each trial of the IAT. In the LMMs, the error term is supposed to follow a normal distribution (i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$), and the dependent variable y is the log transformation of the time response to each trial of the IAT, regardless of whether the answer is correct or not. The expected response y for each observation i ($i \in \{1, \ldots, n\}$) for participant p ($p \in \{1, \ldots, P\}$) on stimulus s ($s \in \{1, \ldots, S\}$) in condition c ($c \in \{1, \ldots, C\}$) can hence be either the *log-odds* of the probability of a correct response (GLMMs) or the log-time of the response (LMMs).

In both GLMMs and LMMs, the fixed intercept α is set at 0. The IAT associative conditions c are specified as the fixed effect $\beta_c X_c$. Since the intercept is set at 0, none of the levels of the fixed effect is taken as reference value. Consequently, the marginal *log-odds* of a correct response for each condition (GLMMs) and the marginal average log-time for each condition (LMMs) are estimated. The fixed part of the models is kept constant, only the random structures change across models.

The GLMMs applied on accuracy responses are identified by a capital "A". The LMMs applied on log-time responses are identified by a capital "T".

The `R` code that can be used for estimating the Rasch model and the log-nromal estimates

from IAT data is illustrated in Appendix A.

4.4.1 Generalized Linear Mixed-Effects Models

Model A1 presents the simplest random structure, where only the between–respondents across–conditions variability and the between–stimuli across–conditions variability are considered by specifying both respondents and stimuli as random intercepts across associative conditions:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i), \tag{4.16}$$

with

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}) \text{ and } \alpha_s \sim \mathcal{N}(0, \sigma^2_{\alpha_s}). \tag{4.17}$$

The random structure of Model A1 results in the estimation of overall respondents' ability estimates θ_p and overall stimuli easiness estimates b_s. This model should be preferred when a low within–respondents between–conditions variability, as well as a low within–stimuli between–conditions variability, are observed. The lack of variability at the levels of both the respondents and the stimuli might indicate a lack of the IAT effect at both levels.

Respondents' ability estimates inform about the overall ability of the respondents in performing the categorization task, and they can be used as a measure of individual differences for further analysis. Stimuli overall easiness estimates provide information on the stimuli functioning in respect to their own category. Stimuli belonging to the same category are supposed to be prototypical exemplars of their own category, and, as such, to be easily recognized and correctly assigned to their category. Consequently, they should have similar easiness estimates. If a stimulus is not recognized as a prototypical exemplar of its alleged category, it will have a higher chance of getting incorrect responses (i.e., being assigned to the incorrect category), from which a lower easiness estimate follows. By comparing the easiness estimates of the stimuli belonging to the same category, it is possible to investigate whether the stimuli belonging to the same category are all easily recognizable as prototypical

exemplars or not.

Model A2 accounts for the within–stimuli between–conditions variability and the between–respondents across–conditions variability. Stimuli are specified as random slopes in the associative conditions, respondents are specified as random intercepts across associative conditions, as follows:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{p[i]} + \beta_{s[i]} c_i + \varepsilon_i), \tag{4.18}$$

with:

$$\beta_{sc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{sc}) \tag{4.19}$$

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}), \tag{4.20}$$

where $\mathbf{\Sigma}_{sc}$ represents the variance-covariance matrix of the population of the stimuli. It expresses the by-stimulus variability in the associative conditions. The higher the covariance of the stimuli in the two conditions, the more similar is their functioning in the two conditions. Model A2 results in condition–specific stimuli easiness estimates b_{sc} and overall ability estimates θ_p. This model would be the the best fitting model when a high within–stimuli between–conditions variability is observed and respondents have a low between–conditions variability.

The low variability at the respondents' level might already indicate a lack of the IAT effect on their accuracy performance (i.e., ability remains constant across conditions). In other words, the IAT associative condition does not have an effect on the respondents' ability to sort the stimuli. As for the overall ability estimates obtained with Model A1, these estimates can be used as a measure of individual differences in performing the categorization task for further analysis. Conversely, the high within–stimuli between–conditions variability indicate that the stimuli functioning is in some way affected by the specific associative condition and that stimuli characteristics (i.e., the category to which they belong) make them more easily categorizable in one condition than in the opposite one. Thus, condition–specific stimuli easi-

ness estimates allow for investigating whether stimuli functioning differs between conditions.

Consider a stimulus representing a can of coke in a Coke-Pepsi IAT. If the stimulus presents a higher easiness estimate in the Coke-Good/Pepsi-Bad condition than in the opposite one, it implies that it was more easily sorted when it shared the response key with *Good* rather than *Bad* attributes. Consequently, the differential measures computed on the condition–specific stimuli estimates inform about the contribution of each stimulus to the IAT effect, which in turn leads to a better understanding of the automatic associations driving the effect.

The random structure of Model A3 has the same level of complexity as that of Model A2. However, the multidimensionality on the error term is specified for the respondents and not for the stimuli. Model A3 accounts for the within–respondents between–conditions variability and between–stimuli across–conditions variability by specifying respondents as random slopes in the associative conditions and stimuli as random intercepts across associative conditions:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{s[i]} + \beta_{p[i]} c_i + \epsilon_i), \tag{4.21}$$

with:

$$\beta_{pc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pc}) \tag{4.22}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{4.23}$$

where $\mathbf{\Sigma}_{pc}$ represents the variance-covariance matrix of the population of the respondents. It expresses the by-respondent variability according to the associative conditions. The high covariance does not necessarily implies that the performance is not affected by the associative condition. For instance, a respondent with a high ability might have a high ability in both conditions, although his performance might be affected by the associative conditions. Model A3 results in condition–specific respondents' ability estimates θ_{pc} and overall easiness estimates b_s. This model would be the best fitting model when a low within–stimuli

between–conditions variability and a high within–respondents between–conditions variability are observed.

As in Model A1, the lack of within–stimuli between–conditions variability might indicate that the stimuli functioning is not affected by the associative condition in which they are presented. The overall easiness estimates can still inform about the stimuli functioning in respect to their own category.

The high within–respondents between–conditions variability at the respondents level indicate that the IAT associative conditions affect the accuracy performance of the respondents, or, in other words, that their ability level is in some way hindered by one of the associative conditions. A measure of the bias due to the associative conditions can be obtained by computing the difference between each respondent condition–specific ability estimate.

4.4.2 Linear Mixed-Effects Models

Model T1 presents the simplest random structure. Only the between–respondents across-conditions variability and the between–stimuli across–conditions variability are considered by specifying both respondents and stimuli as random intercepts across associative conditions:

$$y_i = \alpha + \beta_c X_c + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i, \tag{4.24}$$

with

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}) \text{ and } \alpha_s \sim \mathcal{N}(0, \sigma^2_{\alpha_s}) \tag{4.25}$$

Model T1 allows for estimating overall respondents' speed estimates τ_p and overall stimuli time intensity estimates δ_s. Respondents' speed estimates inform about the overall speed with which they have performed the categorization task. As the ability estimates obtained from Model A1, the overall speed estimates can be used as a measure of individual differences in further analysis. This model should be preferred when a low within–respondents between–

conditions variability and a low within–stimuli between–conditions variability are observed. The lack of variability at both respondents and stimuli levels might indicate that there is no IAT effect at both levels.

As overall easiness estimates, stimuli overall time intensity estimates inform about the stimuli functioning in respect to their own category. If the stimuli belonging to the same category are equally recognized as prototypical exemplars of their own category, they should require a similar amount of time for getting a response, and hence they should have a similar time intensity estimate. If a stimu presents characteristics that make it less recognizable as a prototypical exemplars of a specific category (e.g., a picture of a can of soda that is not immediately recognizable as either Coke or Pepsi), it might require more time for being identified and sorted. Consequently, it should have a higher time intensity estimate. By comparing the time intensity estimates of the stimuli belonging to the same category, it is possible to investigate whether the stimuli belonging to the same category require a similar time for getting a response. In doing so, other stimuli characteristics should be taken into account. For instance, images stimuli require less time to be processed than attribute stimuli (e.g., Houwer & Hermans, 1994). Moreover, the familiarity with a specific term might play an import role in its recognition and sorting, hence positively (if it is a familiar term) or negatively (if it is an unfamiliar term) affecting its time intensity. Also the length of the word itself might influence stimuli time intensity.

Model T2 accounts for the within–stimuli between–conditions variability and the between–respondents across–conditions variability. The random slopes of the stimuli in the associative conditions and the random intercepts of the respondents across associative conditions are specified:

$$y_i = \alpha + \beta_c X_c + \alpha_{p[i]} + \beta_{s[i]} c_i + \varepsilon_i, \tag{4.26}$$

with:

$$\beta_{sc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{sc}) \tag{4.27}$$

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}), \tag{4.28}$$

where $\mathbf{\Sigma}_{sc}$ represents the variance-covariance matrix of the population of the stimuli, and it expresses the by-stimuli variation according to the associative condition. As for accuracy models, the higher the covariance, the more similar the stimuli functioning in the two conditions. Model T2 results in condition–specific stimuli time intensity estimates δ_{sc} and overall speed estimates τ_p. This model should result as the best fitting model when a high within–stimuli between–conditions variability is observed and respondents have a low between–conditions variability.

The low variability at the respondents' level might already indicate a lack of the IAT effect on their speed performance (i.e., speed remains the same across conditions). In other words, the speed of the respondents does not change according to the specific associative condition. As for the overall speed estimates obtained with Model T1, these estimates can be used as a measure of individual differences in performing the categorization task for further analysis.

Conversely, the high within–stimuli between–conditions variability indicate that the stimuli do require a different amount of time to be sorted according to the associative condition in which they are presented. Their functioning is hence affected by the associative conditions, and the condition–specific time intensity allow for investigating how and how much. The differential measure computed between the condition–specific time intensity estimates provide a measure of the bias on the time each stimulus require for getting a response due to the associative conditions. Consequently, the contribution of each stimulus to the IAT effect can be investigated.

The random structure of Model T3 has the same level of complexity as that of Model T2. However, the multidimensionality on the error term is specified for the respondents and not for the stimuli. Model T3 accounts for the within–respondents between–conditions variability and the between–stimuli across–conditions variability by specifying the respondents as random slopes in the associative conditions and the stimuli as random intercepts across

associative conditions:

$$y_i = \alpha + \beta_c X_c + \alpha_{k[i]} + \beta_{j[i]} l_i + \varepsilon_i, \tag{4.29}$$

with:

$$\beta_{pc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pc}), \tag{4.30}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{4.31}$$

where $\mathbf{\Sigma}_{pc}$ is the variance-covariance matrix of the population of the respondents, and it expresses the by-respondents variability according to the associative condition. Similarly to accuracy models, a high covariance does not imply that respondents' performance is not affected by the associative conditions but that their baseline speed is making them having a similar performance in both conditions. This model results in condition–specific respondents' speed estimates τ_{pc} and overall time intensity estimates δ_s. Model T3 should result as the best fitting model when a low within–stimuli between–conditions variability and a high within–respondents between–conditions variability are observed. As in Model T1, the lack of within–stimuli between–conditions variability might indicate that the stimuli functioning is not affected by the associative condition in which they are presented. The overall time intensity estimates can still inform about the stimuli functioning in respect to their own category.

The high within–respondents between–conditions variability at the respondents level indicates that the IAT associative conditions affect the speed performance of the respondents, or, in other words, that their speed is lower in one of the associative conditions. A measure of the bias due to the associative conditions can be obtained by computing the difference between each respondent condition–specific speed estimate.

4.5 Other random structures

The random structures presented in the previous sections are just some of the possible random structures that can be specified for analyzing IAT data. Indeed, since IAT data has a specific structure (illustrated in Section 1.4), a model with a random structure that decomposes error variance into each of the sources of variation can be specified (Maximal Model, MM; Barr et al., 2013).

In the MM, both between–respondents across–conditions variability and within–respondents between–conditions variability is accounted for by specifying respondents random intercepts across conditions and their random slopes in the associative conditions. The same can be done for the stimuli, so that they are specified as random intercepts across conditions and as random slopes in the associative conditions. Moreover, the variability due to the interaction between the stimuli and the respondents variability (i.e., respondents' individual reactions to each stimulus) can be accounted for by specifying the interaction effect between respondents and stimuli random intercepts.

The MM results in the estimation of the weights associated with each level of the fixed effect, as well as in the estimation of the variance of the population to which each factor considered as random belongs. In this case, the stimuli, the respondents, and their interaction. This interaction can be considered as the variability due to the idiosyncratic reactions of each respondent to each stimulus. Also the variance-covariance matrix for each level on which the multidimensionality of the error variance is allowed are estimated. Therefore, the variance of the respondents in each level of the associative condition variable, as well as their covariance, are estimated. The same is done for the stimuli.

By considering the two levels of the fixed effect of the associative conditions and removing the intercept by setting it at 0, this model results in the estimation of 18 parameters, two of which are the weights of the fixed effects. Three parameters refer to the estimated variances of the population of the respondents, that of the stimuli, and the interaction between them. Three parameters are estimated for the multidimensionality of the associative conditions on the respondents (the variance in the two conditions and their covariance), as well

as three parameters for the multidimensionality of the associative conditions on the stimuli (the variance in the two conditions and their covariance). Finally, one parameter refers to the estimated residual variance.

A model of such a complexity needs an extremely high variability at each level of the random structure to converge. Beyond being at risk of convergence failure, it is also at risk of over-fitting the data (Bates, Kliegl, Vasishth, & Baayen, 2015), hence resulting in biased and not-interpretable estimates.

A model with the random structure of the MM is neither needed nor appropriate for the estimation of the Rasch model and the log-normal model estimates from IAT data. By specifying both respondents and stimuli as random intercepts and random slopes in the associative conditions, overall and condition–specific estimates can be obtained for each factor. The difference between each of the condition specific estimates and the overall estimates provides information about the bias due to each condition on either the respondents or the stimuli. The difference between condition–specific estimates results in a measure of the bias due to the IAT associative conditions. Consequently, it allows for investigating the impact of the IAT associative conditions on either respondents' performance or stimuli functioning. When the IAT is used, the focus is usually on this difference, expressing the IAT effect. Therefore, the estimation of the overall estimates for both the respondents and the stimuli can be dropped without losing important information.

For the Rasch model or the log-normal model to be identified, either respondents or stimuli have to be centered around 0 (e.g., Gelman & Hill, 2007). This can be done by setting the fixed intercept at 0 and by specifying either respondents or stimuli as random variation (i.e., random intercepts) around it. As such, each respondents or stimuli BLUP defines the deviation of each level of the considered factor from 0, that is, the average of the respondents or the stimuli estimates. Consequently, only either respondents or stimuli can be specified as random slopes in the associative conditions, while the other must be specified as random intercepts. The decision on where to allow for the multidimensionality of the associative condition, whether on the respondents or on the stimuli, should be driven by the observed variability in the data, also according to the hypotheses og the researcher.

Finally, the estimation of the interaction effect between stimuli and respondents random intercepts does require an high respondents $\times$ stimuli variability to avoid convergence failure. Consequently, it can be dropped and added to the model only in those cases in which the error variance is still high after the estimation of all the other parameters (Judd et al., 2012; Westfall et al., 2014).

Clearly, also other fixed effects could have been included in the model. For instance, the belonging category of the stimuli, which is indeed an independent variable as illustrated in Section 1.4, could have been included as a fixed effect. However, we decided to focus on the effect of the IAT associative condition, and on the deviation from it of each of the levels of the stimuli or the respondents. In our opinion, the information yielded from a model with this structure is more useful in gaining insights on the IAT functioning, for example by highlighting the stimuli giving the highest contribution to the IAT effect. Indeed, by specifying the fixed effect of the stimuli categories, an information on the respondents' or the stimuli deviations from their mean could have been obtained. However, while this information is useful and meaningful for the stimuli functioning, it is not so for the respondents. What does it mean that the respondent p has an impairment of 1.06 on stimulus *pain* of *Bad* category? This information might be more useful if also the interaction between the stimuli categories and the associative conditions is specified. However, this interaction would need an extremely high variability for the model to converge and to provide meaningful estimates.

We decided to keep a more parsimonious model by including as a fixed effect a factor that would have provided useful information regarding both respondents and stimuli, namely, the associative condition. Nothing is preventing anyone for including other fixed effects, and to check whether the model does converge or not. Since the aim of the thesis was to provide a general modeling framework for implicit measures data, we decided to go for a more parsimonious but generalizable model.

Finally, considering only the fixed effect of the condition, hence allowing for the multidimensionality only according to this effect, is in line with previous applications of the Rasch model to IAT data (e.g., Anselmi et al., 2013).

Chapter 5

Applications of (G)LMMs to IAT data

In this Chapter, two empirical applications of the modeling framework proposed in Chapter 4 are presented. In the first application, the accuracy and and log-time models for the estimation of the Rasch model and the log-normal model estimates, respectively, have been applied to an IAT for the implicit assessment of attitudes towards Black and White people (i.e., Race IAT, Section 5.1). The relationship between model estimates and the typical IAT scoring (i.e., *D* score) has been investigated as well. The second application was aimed at investigating whether the estimates obtained with accuracy and log-time models result in a better inference of the construct under investigation than that provided by the *D* score. To pursue this aim, the predictive ability of the model estimates and that of the *D* score have been compared, and an IAT for the implicit assessment of the preference for Dark and Milk chocolate was used (i.e., Chocolate IAT, Section 5.2).

The accuracy and the log-time models were fitted with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in `R` (Version 3.5.1, R Core Team, 2018). The IAT *D* scores were computed by using the `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020c).

A summary of the Rasch model estimates and the log-normal model estimates that can be obtained from the random structures of the (G)LMMs presented in Chapter 4 is reported in Table 5.1

Table 5.1: Rasch model and log-normal model estimates.

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

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

5.1 Empirical application on a Race IAT

5.1.1 Method

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

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

and fast as they could.

5.1.2 Data analysis

Data cleaning and *D* score

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

Outfit Statistics

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

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

The standardized residuals are computed as:

$$z_i = \frac{x_i - P(x_i)}{\sqrt{P(x_i = 1)P(x_i = 0)}}, \tag{5.1}$$

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

Normally, the Outfit statistics are computed by averaging the squared standardized residuals across respondents (stimuli Outfit) or across items (respondents Outfit), and one value for each stimulus and one for each respondent are obtained. In the IAT case, also the associative condition must be taken into account, and the number of Outfit statistics for each respondent and each stimulus depends on the random structure of the model. Let $l \in \{1, \ldots, L\}$ be the number of trials of stimulus s, and assume that each stimulus has an equal number of trials. If this is not true, the variable l takes on the form of l_s.

If Model A2 results as the best fitting model, then condition–specific Outfit statistics u_{sc} for the stimuli:

$$u_{sc} = \frac{\sum_{p=1}^{P}\sum_{l=1}^{L} z_{pscl}^2}{P \times L}, \tag{5.2}$$

and overall Outfit statistics for the respondents u_p:

$$u_p = \frac{\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{c=1}^{C} z_{pscl}^2}{S \times L \times C}, \tag{5.3}$$

are obtained.

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

$$u_{pc} = \frac{\sum_{s=1}^{S}\sum_{l=1}^{L} z_{pscl}^2}{S \times L}, \tag{5.4}$$

and overall Outfit statistics for the stimuli:

$$u_s = \frac{\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{c=1}^{C} z_{pscl}^2}{P \times L \times C}, \tag{5.5}$$

are obtained.

Log-normal model. A similar procedure is followed for the computation of the Outfit statistics on the log-time responses. The difference for the computation of the residuals z_i is taken between the observed log-time responses to each trial t_i and the expected log-time to each trial $\bar{t}_i$ estimated with the log-normal model (Equation 4.12).

If Model T2 results as the best fitting one, then condition–specific Outfit statistics u_{sc} for the stimuli and overall respondents' Outfit u_p statistics can be computed by following Equations 5.2 and 5.3, respectively.

Conversely, if Model T3 results as the best fitting model, respondents' condition–specific Outfit statistics u_{pc} and overall stimuli outfit statistics u_s can be obtained as in Equations 5.4 and 5.5, respectively.

For ~~both~~ the Outfit statistics computed on accuracy responses and log-time responses, the thresholds indicating *underfit* (i.e., data shows a variability that the model cannot explain) or *overfit* (i.e., data shows less variability than that expected by the model) in Linacre (2002) were used to decide on the goodness of fit of the specific respondent/stimulus to the data. If Outfit statistics ranged between 0.50 and 2.00 (Linacre, 2002), the data were considered to have a good fit to the model. A major weight was given to respondents/stimuli showing underfit, while overfit was not considered as much problematic.

Relationship between model estimates and typical scoring

In case the best fitting model allows for the multidimensionality of the associative conditions on the respondents, differential measures (either *ability-differential* or *speed-differential*) are computed. The relationship between the estimates of the Rasch model, those of the log-

normal model, their potential differential measures, and the typical IAT *D* score are investigated both by computing Person's correlations between the variables and by regressing the linear combination of the respondents' estimates on the *D* score. In case of differential measures, both differential measures and their linear combination are regressed on the *D* score in separate models. This is done to determine the actual weight of each condition–specific estimate on the final *D* score. Backward deletion is used for investigating the predictor(s) that explains the higher amount of variance of the *D* score.

5.1.3 Results

No participants or trials were eliminated grounding on the response time exclusion criteria 5.1.2. Three participants were excluded because of the accuracy deletion criterion (Nosek et al., 2002). The sample was finally composed by 62 participants (F $= 48.39\%$, Age = 24.92 ± 2.11 years).

The overall average response time was 815.06ms (*sd* $= 423.20$, *skewness* $= 3.82$, *kurtosis* $= 33.87$), while the average response time was 667.11ms in the WGBB condition (*sd* $= 294.06$, *skewness* $= 4.64$, *kurtosis* $= 44.60$) and 943.01ms (*sd* = 488.89, *skewness* $= 3.45$, *kurtosis* $= 29.05$) in the BGWB one. After the log-transformation of the response latencies (expressed in second), the overall average response time was -0.29 log-seconds (*sd* $= 0.40$, *skewness* $= 0.72$, *kurtosis* $= 3.88$), the average response time was -0.43 log-seconds in the WGBB condition (*sd* $= 0.31$, *skewness* $= 1.26$, *kurtosis* $= 3.73$), and the average response time was -0.15 log-seconds in the BGWB condition (*sd* $= 0.42$, *skewness* $= 0.24$, *kurtosis* $= 5.09$).

Rasch models

The accuracy models in Table 5.1 were applied to the Race IAT. Concerning AIC, Log-Likelihood, and Deviance, Model A2 (AIC = 3784.43, Log-Likelihood = -1886.21, Deviance = 3722.43) performed better than Model A3 (AIC = 3786.51, Log-Likelihood = -1887.26, Deviance = 3774.51) and Model A1 (AIC = 3785.87, Log-Likelihood = -1888.93,

Deviance = 3777.87). However, the latter one showed the lowest BIC value (3813.53, 3825.91, 3828.00, BIC values for Model A1, A2, and A3, respectively). Model A2 was chosen. This model provided overall participants ability parameters θ_p and condition–specific stimuli easiness parameters (b_{WGBB} and b_{BGWB}). The estimates of the fixed effects of Model A2 indicated a higher probability of correct response in the WGBB condition (*log-odds* $= 3.45$, *SE* $= 0.12$) than in the BGWB condition (*log-odds* $= 2.07$, *SE* $= 0.11$). Between–participants variability was 0.17. Between–stimuli variability in the WGBB condition ($\sigma^2 = 0.08$) was lower than the between–stimuli variability in the BGWB condition ($\sigma^2 = 0.15$). The correlation between stimuli variability in the two conditions was moderate ($r = .34$).

Outfit statistics of the respondents ranged between 0.04 and 1.85 (*M* $= 0.92 \pm 0.33$). Seven respondents showed Outfit statistics below 0.50, but 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 5.2.

Table 5.2: Stimuli condition–specific easiness parameters (b_{sc}) and overall time intensity parameters (δ_s) - Race IAT

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

Note: "wf": White person female face; "wm": White person male face; "bf": Black person female face; "bm": Black person male face; WGBB: White-Good/Black-Bad condition; BGWB: Black-Good/White-Bad condition. Rows are ordered by decreasing values of $b_{\text{WGBB}} - b_{\text{WGBB}}$. The units of the easiness estimates are the *log-odds*, the units of the time intensity estimates are the log-seconds.

Overall, the IAT stimuli tended to be easy stimuli. The stimuli tended to be easier in the WGBB condition ($M = 3.44 \pm 0.16$) than in the BGWB condition ($M = 2.05 \pm 0.33$, $t(39) = 19.89$, $p < .001$, 95% CI $[1.24, 1.53]$). The belonging category of the stimuli was used to predict the difference between the condition–specific easiness estimates. The intercept was removed so that one of the categories was taken as a reference for the others. A significant effect of the catgeory of stimuli was found ($F(4, 24) = 359.87$. $p < .001$). The categories of stimuli that gave the highest contribution to the IAT effect were the evaluative dimensions ($B_{\text{Bad}} = 1.49$, $SE = 0.07$, $t(24) = 21.60$, $p < .001$, and $B_{\text{Good}} = 1.62$, $SE = 0.07$, $t(24) = 23.47$,, $p < .001$). The target objects categories gave a lower contribution to the IAT effect ($B_{\text{Black}} = 1.13$, $SE = 0.08$, $t(24) = 14.18$, $p < .001$, $B_{\text{White}} = 1.18$, $SE = 0.08$, $t(24) = 14.88$, $p < .001$). The stimuli that gave the highest contribution to the IAT effect were *joy* and *happiness* (category *Good*), *evil* and *horrible* (category *Bad*), *wm3* and *wf3* (category *White*), and *bm2* and *bf2* (category *Black*). The stimuli that gave the lowest contribution to the IAT effect were *love* and *glory* (category *Good*), *annoying* and *pain* (category *Bad*), *wf1* and *wm1* (category *White*) and *bm3* and *bf3* (category *Black*).

Log-normal models

The log-time models in Table 5.1 were applied to the Race IAT. Model T2 produced aberrant estimates (i.e., correlation between the stimuli random slopes equal to 1). Model T3 (AIC = 4399.66, BIC = 4448.06, Log-Likelihood = -2192.83, Deviance = 4385.66) performed better than Model T1 (AIC = 4762.63, BIC = 4797.20, Log-Likelihood = -2376.32, Deviance = 4752.63). Model T3 was chosen. This model provided condition–specific participants' speed parameters (τ_{WGBB} and τ_{BGWB}) and overall stimuli time intensity parameters δ_j. Respondents' Outfit statistics showed a good fit for all respondents in both associative conditions ($M = 0.98 \pm 0.01$, $Min = 0.98$, $Max = 0.99$ for the BGWB condition, and $M = 0.99 \pm 0.01$, $Min = 0.98$, $Max = 1.03$ for the WGBB condition). Overall Outfit statistics indicated a good fit for all stimuli ($M = 1.00 \pm 0.16$, $Min = 0.77$, $Max = 1.33$).

Responses in the WGBB condition were faster ($B = -0.43$, $SE = 0.02$) than responses in the BGWB condition ($B = -0.15$, $SE = 0.03$). The between-stimuli variability was

particularly low ($\sigma^2 = 0.003$), while the between–participants variability was slightly higher in the BGWB condition ($\sigma^2 = 0.05$) than in the WGBB one ($\sigma^2 = 0.02$). The correlation between respondents' variability in the two conditions was strong ($r = .63$).

Stimuli time intensity parameters δ_s obtained from Model T3 are reported in Table 5.2. A significant effect of the belonging categories of the stimuli was found on the time intensity estimates ($F(4, 2$ = 11.77, $p < .001$). The exemplars of both evaluative dimensions tended to require a high amount of time for getting a response ($B_{\text{Bad}} = 0.04$, $SE = 0.01$, $t(24) = 3.44$, $p < .001$, and $B_{\text{Good}} = 0.03$, $SE = 0.01$, $t(24) = 2.63$, $p = .01$). The exemplars of the category *Black* were the stimuli requiring the least time for getting a response ($B = -0.07$, $SE = 0.01$, $t(24) = -4.88$, $p = 0.01$), immediately followed by the exemplars of the category *White* ($B = -0.03$, $SE = 0.01$, $t(24) = -2.13$, $p = 0.04$).

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

Relationship between model estimates and typical scoring

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

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

negatively predicted the *D* score ($B = -0.18$, $t(59) = -2.48$, $p = .016$).

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

5.1.4 Final remarks

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

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

effect in Anselmi et al. (2013).

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

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

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

5.2 Empirical application on a Chocolate IAT

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

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

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

5.2.1 Method

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

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

appropriately modified to represent either *Dark* or *Milk* chocolate.

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

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

5.2.2 Data analysis

Data cleaning and *D* score computation

Exclusion criteria based on both accuracy (Nosek et al., 2002) and time responses (Greenwald et al., 2003) are applied (see Section 5.1.2). The *D4* algorithm in Greenwald et al. (2003) is used to score the IAT. The difference is taken between the average response time in the MGDB condition and that in the DGMB condition. Therefore, positive scores stand for a possible preference for Dark chocolate over Milk chocolate. LMMs are applied to the raw log-time responses of both correct and incorrect responses, without any penalties.

Oufit statistics

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

Relationship between model estimates, typical scoring, and explicit measures

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

Predictive ability of a behavioral outcome

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

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

Predictive ability of the reduced data sets. The information at the stimuli level can be used to select the most and least informative stimuli for each stimuli categories. The two most informative stimuli for each category, and the least informative stimuli for each category, are selected to create smaller data sets, an highly informative one ("Best") and a lowly informative one ("Worst"). In both cases, the stimuli pool is composed of 8 stimuli. A *D* score is computed on each of the newly obtained data set, and it is used for predicting the choice. Their performance is compared with that of the *D* score computed on the entire data

set. All starting models include food habits, and relevant predictors are selected with backward deletion. Model general accuracy (i.e., percentage of choices correctly identified by the model), model DCC accuracy (i.e., percentage of DCCs correctly identified by the model), and model MCC accuracy (i.e., percentage of MCCs correctly identified by the model) are used as criteria to establish the predictors best accounting for the actual choice. Nagelkerke's R^2 (Nagelkerke, 1991) is used as Pseudo R^2.

5.2.3 Results

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

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

Rasch models

The accuracy models presented Table 5.1 were applied to the Chocolate IAT. Model A3 failed to converge, while Model A2 (AIC = 3625.58, Log-Likelihood = -1806.79, Deviance = 3613.58) performed better than Model A1 (AIC = 3627.71, Log-Likelihood = -1809.85, Deviance = 3619.71). Model A1 showed a lower value of BIC than Model A2 (3656.07, 3668.13 for Model A1 and Model A2, respectively). Model A2 was chosen. The model resulted in the estimation of overall respondents' ability θ_p and condition–specific easiness

parameters (b_{MGDB} and b_{DGMB}).

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

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

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

Stimuli easiness parameters are reported in Table 5.3.

Table 5.3: Stimuli condition–specific easiness parameters (b_{sc}) and overall time intensity parameters (δ_s) - Chocolate IAT

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

Note: DGMB: Dark-Good/Milk-Bad condition; MGDB: Milk-Good/Dark-Bad condition; Difference: Difference be[illegible]n DGMB and MGDB condition. Rows are ordered by absolute decreasing values of $b_{\text{DGMB}} - b_{\text{DGMB}}$. The units of the easiness estimates are the *log-odds*, the units of the time intensity estimates are the log-seconds.

Irrespective of the category to which they belong, stimuli tended to be easier in the MGDB condition ($M = 3.63 \pm 0.29$) than in the DGMB one ($M = 2.60 \pm 0.40$) $= -21.97$. $p < .001$, 95% CI $[-1.13, -0.94]$). A significant effect of the categories of the stimuli was found on the difference in the easiness estimates between the associative conditions. The exemplars of the category *Milk* ($B = -1.13$, *SE* $= 0.11$, $t(36) = -10.84$, $p < .001$) and those of the category *Good* ($B = -1.09$, *SE* $= 0.08$, $t(36) = -14.10$, $p < .001$) were the stimuli that gave the highest contribution to the IAT effect. The category of stimuli that gave the least contribution to the IAT effect was the category *Bad* ($B = -0.92$, *SE* $= 0.07$, $t(36) = -11.98$, $p < .001$), followed by the contribution given by the category *Dark* ($B = -1.05$, *SE* $= 0.11$, $t(36) = -9.97$, $p < .001$).

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

Log-normal models

The log-time models presented Table 5.1 were applied to the Chocolate IAT. Model T2 produced aberrant estimates. Model T3 (AIC = 7159.23, BIC = 7208.87, Log-Likelihood = -3572.62, Deviance = 7145.23) performed better than model T1 (AIC = 7856.45, BIC = 7891.91, Log-Likelihood = -3923.23, Deviance = 7846.45). Thus, model T3 was chosen. The model resulted in overall stimuli time intensity parameters δ_k and respondents' condition–specific speed parameters (τ_{MGDB} and τ_{MGDB}). Responses tended to be faster in the MGDB condition ($B = -0.36$, *SE* $= 0.02$) than in the DGMB condition ($B = -0.12$, *SE* $= 0.03$). Between-stimuli variability was extremely low ($\sigma^2 = 0.004$). Between-participants variability was higher in the DGMB condition ($\sigma^2 = 0.05$) than in the MGDB one ($\sigma^2 = 0.03$). The correlation between participants' slopes in the two conditions was moderate

($r = .40$).

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

Stimuli time intensity parameters δ_k are reported in Table 5.3. A significant effect of the categories of the stimuli was found on the stimuli time intensity estimate ($F(4, 36) = 37.41$, $p < .001$). The exemplars of both the target objects categories required the least amount of time for getting a response ($B_{\text{Dark}} = -0.09$, $SE = 0.01$, $t(36) = -8.99$, $p < .001$, and $B_{\text{Milk}} = -0.04$, $SE = 0.01$, $t(36) = -4.09$, $p < .001$). The exemplars of the category *Bad* were the stimuli that required the highest amount of time for getting a response ($B = 0.05$, $SE = 0.01$, $t(36) = 6.20$, $p < .001$), followed by those belonging to the category *Good* ($B = 0.03$, $SE = 0.01$, $t(36) = 3.70$,$p < .001$). It was possible to identify stimuli with time intensity estimates far away from the time intensity estimates of the stimuli belonging to the same category. For instance, stimulus *heaven* was the stimulus requiring more time within the *Good* category.

Relationship between model estimates, typical scoring, and explicit measures

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

Results of Pearson's correlations computed between respondents' explicit preference for Milk and Dark chocolate, *D* scores, ability estimates, condition–specific speed estimates, and *speed-differential* are reported in Table 5.4.

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

Table 5.4: Correlation between model estimates, explicit measures, and *D* scores.

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

Note: $^{***}\ p < .001$, $^{**}\ p < .01$, *D*: IAT *D* score, τ: speed estimate, θ: Ability estimate, DGMB: Dark-Good/Milk-Bad condition, MGDB: Milk-Good/Dark-Bad condition, *Speed-differential*: $\tau_{\text{MGDB}} - \tau_{\text{DGMB}}$.

was associated with *Good* attributes, as it is pointed out by the negative correlations between explicit evaluations and condition–specific speed estimates.

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

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

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

Predictive ability of a behavioral outcome

The information provided by the difference between the condition–specific easiness estimates (see comment to Table 5.3 in Section 5.2.3) was used to create two smaller data sets. The starting data set was composed of 8,879 observations. In a first data set (“Best”), only the

responses to the stimuli giving the highest contribution to the IAT effect were selected. This data set resulted in a total of 2,941 observations, ranging from a minimum of 38 observations to a maximum of 41 observations per participant. In a second data set ("Worst"), only the responses to the stimuli giving the lowest contribution to the IAT effect were selected. This data set resulted in a total number of 2,587 observations, with a minimum of 38 observations and a maximum of 42 observations per participant. The *D4* algorithm was computed for both the Best data set and the Worst data set. In both cases, the data set was reduced to about $1/3$ of the total number of observations.

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

Results of backward deletion are reported in Table 5.5. The *Speed-differential* showed a slightly better general accuracy, due to a small gain in DCC accuracy, than the *D* score. Interestingly, the *D* scores computed on both Best and Worst data sets showed a better general accuracy than both the *D* score computed on the entire data set and the *speed-differential*. The *D* score computed on the Best data set showed the highest general accuracy, resulting from a gain on both DCC accuracy and MCC accuracy. It also explained the highest proportion of variance. Conversely, the *D* score computed on the Worst data set explained the lowest proportion of variance.

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

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

Table 5.5: Choice prediction results for the differential measures and their Single components.

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

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

5.2.4 Final remarks

Results of the study reported in this section corroborate the higher reliability of the estimates obtained with statistical models able to account for IAT error variance and its random struc-

ture than that of the typical scoring methods of the IAT. The information at the stimuli level obtained can be used to reduce the across-trial variability, hence resulting in a better IAT measure as expressed by the *D* score.

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

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

Also the overall time intensity estimates provided useful information on the stimuli functioning. Time intensity estimates highlighted different processing times both between stimuli categories (i.e., images require less time for getting a response than attributes) and within the same stimuli category (i.e., the stimuli showing a time intensity estimate far away from the estimates of the stimuli belonging to the same category).

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

ing out the respondent. In an online setting, a shorter administration time is definitely a good incentive for the participation, and might prevent respondents' withdrawal out of boredom and/or tiredness.

Chapter 6

Multiple implicit measures: Models specification

As already illustrated in Chapter 2, the IAT and the SC-IAT can be administered together for obtaining both a comparative measure of the preference for one target object over the other and an absolute evaluation towards each of them. Their data are usually analyzed separately, and separate *D* scores are computed for each measure, following an approach such as that in Chapter **??**. By doing so, neither the sources of variability due the fully-crossed structure of each implicit measure nor those due to the presentation of the implicit measures to the same respondents are accounted for. Additionally, no attempts of modeling the SC-IAT data within a Rasch framework or of having a comprehensive model for the IAT and the SC-IAT have been made so far. In this chapter, we aim to fill these gaps by introducing a comprehensive modeling approach for multiple implicit measures (i.e., the IAT and the SC-IAT). This modeling framework is obtained by exploiting the flexibility of Linear Mixed-Effects Models (LMMs) for obtaining the estimates of the Rasch model and the log-normal model parameters.

Two levels of model complexity are presented. At a first level, each implicit measure is modeled separately by employing the models presented in Chapter 4. These models will not be further illustrated here. Only a brief summary of the Rasch model and the log-normal

model parameters that can be estimated is provided. At a second level, the between–measures variability is accounted for by considering the within–respondents between–measures variability (Model 2) or the within–respondents between–conditions variability across implicit measures (Model 3).

6.1 Single measures models

The models presented in Chapter 4 for both accuracy and log-time responses can be used for modeling each implicit measure separately. This implies that, even though the variability within each measure is accounted for, the between–measures variability at both the respondents and the stimuli levels can still affect the parameter estimates. Nonetheless, this approach is valid when implicit measures are administered as stand alone measures.

In this section, the IAT and the SC-IAT data have been analyzed separately mainly for two reasons: (i) to investigate whether the modeling approach for IAT data in Chapter 4 can be extended to SC-IAT data, and (ii) to investigate whether and how these estimates are different from the ones obtained with a more sound approach accounting for the between–measures sources of variability.

Irrespective of the implicit measure or the dependent variable (i.e., accuracy or log-time responses) under investigation, the fixed intercept was set at 0. The effect of the associative condition of each implicit measure was the fixed effect in all models. Since the fixed intercept was set at 0, the fixed effects can be interpreted as the estimates of either the expected *log-odds* of the probability of a correct response in each associative condition (Accuracy models) or the expected average log-time responses in each associative condition (Log-time models).

Model 1 (the Null model) accounts for the between–respondents and between–stimuli variability across associative conditions by specifying them as random intercepts. For each separate implicit measure ($m \in \{1, \ldots, M\}$, where m is the number of implicit measures), overall estimates at the respondents' (θ_{pm} and τ_{pm}) and the stimuli (b_{sm} and δ_{sm}) levels can be obtained for the Rasch and the log-normal models.

The random structure of Model 2 results in the estimation of condition–specific stimuli

parameters (b_{scm} and δ_{scm}) and overall respondents' parameters (θ_{pm} and τ_{pm}) for each implicit measure by considering the random slopes of the stimuli in the associative conditions and the random intercepts of the respondents across conditions. This model accounts for the within–stimuli between–conditions variability and the between–respondents variability across conditions.

Finally, the random structure of Model 3 results in the estimation of overall stimuli parameters (b_{sm} and δ_{sm}) and condition–specific respondents' parameters (θ_{pcm} and τ_{pcm}) are obtained by considering the random intercepts of the stimuli across conditions and the random slopes of the respondents in the associative conditions. This model accounts for the between–stimuli across conditions variability and the within–respondents between–conditions variability.

By separately analyzing the data obtained from implicit measures originally administered together, the within–respondents between–measures variability, as well as the within–stimuli between–measures variability, are neglected. Therefore, the estimates of the models parameters can still be affected by error variance components. Moreover, since the estimates for both the respondents and the stimuli are obtained from separate, independent models, they cannot be directly compared between each other.

The models presented in Section 6.2 overcome this issue by considering data from different implicit measures altogether.

6.2 Comprehensive models

The models presented in this section are identified by the superscript "C" (i.e., "Comprehensive"). Data from the IAT and the SC-IATs are considered and modeled together. In all models, the fixed intercept is set at 0, while the fixed effect varies. Specifically, in the Null model and in Model 2^{C}, the fixed effect β is the type of implicit measure, while in Model 3^{C} the fixed effect β is the effect of the associative condition of each implicit measure.

The only differences concerning the models applied on accuracy or log-time responses concern the dependent variable and the assumption on the distribution of the error terms.

GLMMs (Section 6.2.1) are applied on accuracy responses, and the error term ε_i is assumed to follow a logistic distribution. These models are identified with a capital "A". LMMs (Section 6.2.2) are applied on log-time responses and the error term ε_i is assumed to follow a normal distribution. These models are identified with a capital "T".

In all models, only the between–stimuli across–measures variability is considered. The investigation of the stimuli functioning according to the specific implicit measure would indeed provide interesting information. For instance, the SC-IAT is known to be an easier task than the IAT. By having an information at the stimuli level, it would be possible to understand whether only some stimuli contribute to make the task easier. Nonetheless, for specifying the random slopes of the stimuli in each implicit measure, a high within–stimuli between–measures variability is needed, but previous studies (e.g., Epifania, Robusto, & Anselmi, 2020c, 2020a) already highlighted a low within–stimuli between–conditions variability, especially for what concerns time responses. Moreover, the focus is more oriented on understanding the intra- and inter-individuals differences in performing at different implicit measures. Consequently, multidimensionality of the error variance was allowed only at the level of the respondents.

6.2.1 Comprehensive GLMMs

Model A1$^{\text{C}}$ is considered as the Null model:

$$y_i = logit^{-1}(\alpha + \beta_m X_m + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i), \tag{6.1}$$

with

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}) \text{ and } \alpha_s \sim \mathcal{N}(0, \sigma^2_{\alpha_s}), \tag{6.2}$$

where both respondents' and stimuli are specified as random intercepts across associative conditions and type of implicit measure.

Model A1$^{\text{C}}$ results in overall respondents ability estimates (θ^{C}_p) and overall stimuli eas-

iness estimates (b_s^C) of the Rasch model. These estimates inform about the general ability of the respondents to perform the categorization task and the overall easiness of the stimuli across implicit measures. This model should be preferred when both a low within–respondents and between–measures variability and a low within–stimuli between–measures variability are observed. The lack of variability at both levels might already indicate that respondents' ability is not affected by the specific implicit measure or, in other words, that their ability is constant across measures. Similarly, stimuli easiness does not vary across implicit measures.

In Model A2^C, between–stimuli variability across implicit measures and within–respondents between–measures variability are accounted for by specifying stimuli random intercepts and respondents' random slopes in the implicit measures:

$$y_i = logit^{-1}(\alpha + \beta_m X_m + \alpha_{k[i]} + \beta_{p[i]} m_i + \varepsilon_i), \tag{6.3}$$

with:

$$\beta_{pm} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pm}), \tag{6.4}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{6.5}$$

where $\mathbf{\Sigma}_{pm}$ is the variance-covariance matrix of the population of the respondents and it expresses the by-respondents variability according to the implicit measure. Model A2^C results in overall stimuli easiness estimates across implicit measures (b_s^C) and measure–specific respondents' ability estimates (θ_{pm}^C). A high within–respondents between–measures variability is needed for this model to be the best fitting one. This variability indicates that respondents' ability performance is affected by the specific implicit measure. The estimates provided by this model can hence inform about the change in the ability performance of the respondents in each implicit measure. However, no information on the effect of the associative condition is available.

In both Model A1^{C} and Model A2^{C}, the fixed effect is the type of measure. Therefore, it provides the expected *log-odds* of the probability of a correct response in each implicit measure. Since these estimates are obtained from the same model, they can be directly compared between each other.

The random structure of Model A3^{C} accounts for the between–stimuli variability across implicit measures and the within–respondents between–conditions variability. Stimuli random intercepts and respondents' random slopes in each associative condition of each implicit measure are specified:

$$y_i = logit^{-1}(\alpha + \beta_c X_c + \alpha_{s[i]} + \beta_{p[i]c_i} + \varepsilon_i), \tag{6.6}$$

with:

$$\beta_{pc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pc}), \tag{6.7}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{6.8}$$

where $\mathbf{\Sigma}_{sc}$ is the variance-covariance matrix of the population of the respondents, expressing the by-respondent adjustment in each associative condition on each implicit measure. Model A3^{C} results in overall stimuli easiness estimates (b_s^{C}) and condition–specific respondents' ability estimates, for each implicit measure (θ_{pmc}^{C}). Model A3^{C} requires a high within–respondents between–conditions variability to result as the best fitting model. The high variability between the responses of the participants in each condition of each measure already stands for an effect of the associative conditions determined by each implicit measure on respondents' performance. By taking the difference between the condition–specific estimates of each implicit measure, a measure of the bias on respondents' performance due to the associative conditions can be obtained.

6.2.2 Comprehensive LMMs

Models with the same random structures as those presented in Section 6.2.2 are specified for obtaining the log-normal model estimates from the log-time responses.

Model $T1^C$ accounts for the between–respondents and the between–stimuli variability across implicit measures. As such, it is taken to be as the Null model:

$$y_i = \alpha + \beta_m X_m + \alpha_{p[i]} + \alpha_{s[i]} + \varepsilon_i, \tag{6.9}$$

with

$$\alpha_p \sim \mathcal{N}(0, \sigma^2_{\alpha_p}), \text{ and } \alpha_s \sim \mathcal{N}(0, \sigma^2_{\alpha_s}). \tag{6.10}$$

The random structure specification of Model $T1^C$ results in the estimation of overall respondents' speed parameters (τ_p^C) and overall stimuli time intensity parameters (δ_s^C). Consequently, only general information about the respondents' performance and the stimuli functioning across implicit measures is available.

Model $T2^C$ accounts for the between–stimuli variability across implicit measures and the within–respondents between–measures variability by specifying the random intercepts of the stimuli and the random slopes of the respondents in the implicit measures:

$$y_i = \alpha + \beta_m X_m + \alpha_{s[i]} + \beta_{p[i]} m_i + \varepsilon_i, \tag{6.11}$$

with:

$$\beta_{pm} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pm}), \tag{6.12}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{6.13}$$

where $\mathbf{\Sigma}_{pm}$ represents the variance-covariance of the population of the respondents, express-

ing their variability due the effect of the implicit measure. Model T2$^{\text{C}}$ results in overall stimuli time intensity estimates across implicit measures (δ_s^{C}) and measure–specific respondents' speed estimates (τ_{pm}^{C}). A high within–respondents between–measures variability is needed for this model to be the best fitting one. This variability indicates that respondents' speed is affected by the specific implicit measure. However, it is not possible to rule out the possibility that this variability is due to the effect of the associative conditions.

The fixed effect in both Model T1$^{\text{C}}$ and Model T2$^{\text{C}}$ is the type of implicit measure. Therefore, the expected average log-times in each implicit measure are obtained.

The variability due to the effect of the associative conditions of each implicit measure can be understood with the random structure specification of Model T3$^{\text{C}}$. By specifying the respondents' random slopes in each associative condition of each implicit measure, this model accounts for the within–respondents between–conditions and between–measures variability, as well as the between–stimuli across–measures variability:

$$y_i = \alpha + \beta_c X_c + \alpha_{s[i]} + \beta_{p[i]c} c_i + \varepsilon_i, \tag{6.14}$$

with:

$$\beta_{pc} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{pc}), \tag{6.15}$$

$$\alpha_s \sim \mathcal{N}(0, \alpha_s^2), \tag{6.16}$$

where $\mathbf{\Sigma}_{pc}$ represents the variance-covariance matrix of the population of the respondents, expressing the variability due to their adjustments to each of the associative conditions in each of the implicit measures. Model T3$^{\text{C}}$ results in overall stimuli time intensity estimates (δ_s^{C}) and condition–specific respondents' speed estimates, for each implicit measure (τ_{pmc}^{C}). Model T3$^{\text{C}}$ should be preferred when a high within–respondents between–conditions variability is observed. The high variability between respondents' in each condition of each measure stands for an effect of the associative conditions determined by each implicit mea-

sure on respondents' performance. By taking the difference between the condition–specific estimates of each implicit measure, a measure of the bias on respondents' performance due to the associative conditions can be obtained.

A measure of the bias due to the associative conditions of each implicit measure can be obtained from the estimates provided by the Single measures models in Section 6.1 as well. However, as already stated, those estimates are affected by both the within–respondents between–measures variability and the within–stimuli between–measures variability. Conversely, these sources of variability are accounted for in the comprehensive modeling framework, potentially resulting in more reliable estimates. Moreover, the estimates obtained with the comprehensive modeling approach are directly comparable between each other because they are derived from the same model.

In the next chapter, an empirical application of the comprehensive modeling framework and the separate modeling of each implicit measure is illustrated. The estimates of the Rasch model parameters and those of the log-normal model parameters are used for predicting a behavioral outcome, and their predictive performances are compared with those of the typical scoring method of the IAT and the SC-IAT. The relationship between the model estimates and the typical scores of these implicit measures are investigated as well.

Chapter 7

Multiple implicit measures: Empirical applications

In this chapter, the accuracy and log-time responses of the IAT and the SC-IAT have been analyzed following both a single measures approach, as described in Section 6.1 of Chapter 6, and the comprehensive modeling approach presented in Section 6.2 of Chapter 6.

Table 7.1 summarizes the Rasch model parameters and the log-normal model parameters that can be obtained from the random structures specification in Chapter 6. The specific fixed effects for each model are illustrated as well. The `R` code used for estimating the models is reported in Appendix B

Accuracy and log-time models were fitted with the `lme4` package (Bates, Mächler, et al., 2015) in `R` (Version 3.5.1, R Core Team, 2018) (`bobyqa` Optimizer). The *D* scores of the IAT and the SC-IATs were computed with the `implicitMeasures` package (Epifania, Anselmi, & Robusto, 2020c).

7.1 Method

A Chocolate IAT, a Milk Chocolate SC-IAT, and a Dark Chocolate SC-IAT were used. Data are the same as those in Epifania, Anselmi, and Robusto (2020a), that have been presented in Chapter 2.

Table 7.1: Overview of the accuracy and log-time models.

Model	Fixed effect	Respondents	Stimuli
		Single measures models	
A1a	Associative condition	Overall (θ_{pm})	Overall (b_{sm})
T1a	Associative condition	Overall (τ_{pm})	Overall (δ_{sm})
A2	Associative condition	Overall (θ_{pm})	Condition–specific (b_{scm})
T2	Associative condition	Overall (τ_{pm})	Condition–specific (δ_{scm})
A3	Associative condition	Condition–specific (θ_{pcm})	Overall (b_{sm})
T3	Associative condition	Condition–specific (τ_{pcm})	Overall (δ_s)
		Comprehensive models	
A1b	Implicit measure	Overall (θ_p^{C})	Overall (b_s^{C})
T1b	Implicit measure	Overall (τ_p^{C})	Overall (δ_s^{C})
A4	Implicit measure	Measure–specific (θ_{pm}^{C})	Overall (b_s^{C})
T4	Implicit measure	Measure–specific (τ_{pm}^{C})	Overall (δ_s^{C})
A5	Associative condition	Condition–specific ($\theta_{pcm}^{\mathrm{C}}$)	Overall (b_s^{C})
T5	Associative condition	Condition–specific (τ_{pcm}^{C})	Overall (δ_s^{C})

Note: $p \in \{1, \ldots, P\}$, $s \in \{1, \ldots, S\}$, $c \in \{1, \ldots, C\}$, $m \in \{1, \ldots, M\}$, denote any respondent, stimulus, condition, implicit measure, where P, S, C, and M are the number of respondents, stimuli, conditions, and implicit measures respectively, C: Estimates obtained with a comprehensive modeling of IAT and SC-IAT responses.

For the description of the sample, the stimuli and materials employed please refer to Section 2.3.1 of Chapter 2.

7.1.1 Data analysis

Data cleaning and typical scoring of implicit measures

The *D4* algorithm in Greenwald et al. (2003) was used for computing the IAT *D* score (i.e., trials $>$ 10,000 ms were discarded, error responses were replaced by the average response time inflated by a 600 ms penalty). The difference was taken between the average response time in the Milk/Good-Dark-Bad condition (MGDB) and that in the Dark-Good/Milk-Bad condition (DGMB). Positive scores stand for a possible preference for Dark chocolate over

Milk chocolate.

The procedure in Karpinski and Steinman (2006) was followed for computing the SC-IATs D score (i.e., trials < 350 ms were discarded, error responses were replaced by the average response time inflated by a 450 ms penalty). The difference was computed between the average response time in the condition where the target chocolate was associated with negative attributes and that where it was associated with positive attributes. Positive scores stand for a positive evaluation of the target chocolate.

The raw log-times of both correct and incorrect responses were used for the estimation of the log-normal models. No correction on the incorrect responses was applied.

Relationship between model estimates and typical scoring

The relationship between the Rasch model and the log-normal model estimates and the typical scores of implicit measures are investigated. Both the estimates obtained from the separate modeling of each implicit measure (i.e., single measure models) and those obtained from the comprehensive modeling (i.e., comprehensive model) are used.

Regardless of the dependent variable (either accuracy responses or log-time responses) and the type of modeling (single measure vs comprehensive), if the best fitting model is a model resulting in condition–specific respondents' estimates, differential measures are computed by taking the difference between the condition–specific estimates. The differential measures (i.e., *ability-differential* and/or *speed-differential*) express the bias on respondents' accuracy or speed performance due to the effect of the associative conditions.

The model estimates are used to predict their respective typical scoring for each implicit measure. A stepwise approach with forward selection is followed to select the predictors best accounting for the dependent variable. All full models are compared against the same Null model, which included only the estimation of the intercept. The estimate of the intercept is the expected average of the typical score.

Prediction of a behavioral outcome

The predictive ability of the Rasch model and the log-normal model estimates are compared with that of the typical scoring methods of implicit measures. In case the best fitting model allows for the multidimensionality at the respondents' level, the differential measures (i.e., *ability-differential* and/or *speed-differential*) are used for the prediction of the behavioral outcome as well. Dark chocolate choice (DCC) is labeled as 0 and Milk chocolate choice (MCC) is labeled as 1. The predictive ability of the linear combination of the IAT *D* with the single SC-IAT scores (i.e., *D-Dark* and *D-milk*) and that of the linear combination of the IAT *D* with a differential SC-IAT score (i.e., *D-Sciat*, difference between *D-Dark* and *D-Milk*) are investigated. The linear combinations of the single components of each typical scoring (i.e., the average response time computed on the corrected latencies in each associative conditions) are considered as well for predicting the choice.

A stepwise approach with forward selection is followed, and *Nagelkerke's* R^2 (Nagelkerke, 1991) is computed as a *Pseudo* R^2. To investigate the linear combination of the predictors that best accounts for the choice, model general accuracy (i.e., the ratio between the choices correctly identified by the model and the total number of choices), DCCs accuracy (i.e., the ratio between the DCCs correctly identified by the model and the number of observed DCCs), and MCCs accuracy (i.e., the ratio between the MCCs correctly identified by the model and the number of observed MCCs) are computed.

7.2 Results

Data from nine participants were discarded. Eight of them explicitly reported not understanding the tasks they were asked to perform in either the IAT or one of the SC-IATs. One participant showed too many fast responses, specifically in the Dark SC-IAT (more than 30% of responses with a latency lower than 350 ms) and was removed. The final sample was composed of 152 participants (F $= 63.82\%$, Age = 24.03 ± 2.82). Milk chocolate was chosen by 48.03% of the participants.

The descriptive statistics of the response times for each implicit measure (and their associative conditions) are reported in Section 2.3.3 of Chapter 7. The descriptive statistics of the log-response times are here reported.

In the IAT, the overall average log-response time was -0.26 log-seconds (*sd* $= 0.43$, *skewness* $= 0.90$, *kurtosis* $= 2.34$). The average log-response time in the DGMB condition was -0.14 log-seconds (*sd* $= 0.45$, *skewness* $= 0.56$, *kurtosis* $= 2.72$) and that in the MGDB condition was -0.37 log-seconds (*sd* $= 0.37$, *skewness* $= 1.32$, *kurtosis* $= 2.69$).

The overall average log-response time in the Dark SC-IAT was -0.46 log-seconds (*sd* $= 0.35$, *skewness* $= 1.33$, *kurtosis* $= 3.21$). The average log-response time in the DB condition was -0.47 log-seconds (*sd* $= 0.35$, *skewness* $= 1.31$, *kurtosis* $= 3.17$) and that in the DG condition was -0.45 log-seconds (*sd* $= 0.35$, *skewness* $= 1.36$, *kurtosis* $= 3.25$).

The overall average log-response time in the Milk SC-IAT was -0.46 log-seconds (*sd* $= 0.34$, *skewness* $= 1.18$, *kurtosis* $= 4.03$). The average log-response time in the MB condition was -0.44 log-seconds (*sd* $= 0.34$, *skewness* $= 1.78$, *kurtosis* $= 4.03$) and that in the MG condition was -0.48 log-seconds (*sd* $= 0.33$, *skewness* $= 1.18$, *kurtosis* $= 4.09$).

7.2.1 Single measures models

Rasch models

Model comparison is reported in Table 7.2. Model A3 resulted as the best fitting model for all implicit measures. Condition–specific ability estimates (θ_{DGMB}, θ_{MGDB}, θ_{DG}, θ_{DB}, θ_{MG}, θ_{MB}), and overall stimuli easiness estimates b_{sm} for each implicit measure were obtained.

A higher probability of a correct response was observed in the MGDB condition (*log-odds* $= 4.00$, *SE* $= 0.13$), in the DB condition (*log-odds* $= 3.49$, *SE* $= 0.12$), and in the MG condition (*log-odds* $= 3.49$, *SE* $= 0.11$), than in their respective contrasting ones (*log-odds* $= 2.87$, *SE* $= 0.08$, *log-odds* $= 3.28$, *SE* $= 0.11$, and *log-odds* $= 3.30$, *SE* $= 0.11$, for DGMB, the DG, and the MB conditions, respectively). Respondents' showed a higher variability in the MGDB condition ($\sigma^2 = 1.05$), in the DB condition ($\sigma^2 = 0.83$), and in the MB condition ($\sigma^2 = 0.76$) than in their respective contrasting conditions ($\sigma^2_{\text{DGMB}} = 0.46$,

Table 7.2: Model comparison - Single measures.

	Model	AIC	BIC	Log-Likelihood	Deviance
IAT	A1	6733.40	6764.60	−3362.70	6725.40
	T1	16258.00	16297.00	−8123.90	16248.00
	A2	6719.20	6766.00	−3353.60	6707.20
	T2		Aberrant estimates		
	A3	6631.10	6678.00	−3309.60	6619.10
	T3	14903.00	14957.00	−7444.30	14889.00
Dark SC-IAT	A1	8122.90	8154.90	−4057.40	8114.90
	T1	12160.00	12200.00	−6075.10	12150
	A2	8125.70	8173.70	−4056.90	8113.70
	T2		Aberrant estimates		
	A3	8013.10	8061.10	−4000.60	8001.10
	T3	11973.00	12029.00	−5979.70	11959.00
Milk SC-IAT	A1	8074.50	8106.40	−4033.20	8066.50
	T1	12362.00	12402.00	−6176.20	12352
	A2	8045.30	8093.20	−4016.60	8033.30
	T2		Aberrant estimates		
	A3	7925.20	7973.10	−3956.60	7913.20
	T3	12120.00	12176.00	−6052.80	12106.00

Note: "A": Accuracy Models, "T": Log-time models

$\sigma^2_{\text{DG}} = 0.65$, and $\sigma^2_{\text{MG}} = 0.69$). Variability at the stimuli level was 0.04, 0.17, and 0.16 for the IAT, the Dark SC-IAT, and the Milk SC-IAT, respectively.

The stimuli easiness estimates for the IAT, the Dark SC-IAT, and the Milk SC-IAT are reported in Table 7.3.

Table 7.3: Single measure model: Stimuli easiness estimates (b_{km}) and time intensity estimates (δ_{km}).

	b			δ				b			δ		
	IAT	Dark SC-IAT	Milk SC-IAT	IAT	Dark SC-IAT	Milk SC-IAT		IAT	Dark SC-IAT	Milk SC-IAT	IAT	Dark SC-IAT	Milk SC-IAT
Bad attributes							*Good* attributes						
agony	−0.14	−1.06	−1.03	0.09	0.08	0.08	beautiful	−0.01	−0.36	−0.02	0.01	−0.01	0.01
annoying	−0.30	−0.92	−0.70	0.11	0.12	0.12	excellent	0.12	−0.06	0.03	0.02	0.05	0.07
bad	−0.21	0.04	0.07	0.04	0.01	0.01	glory	−0.18	0.44	0.37	0.04	0.02	0.03
disaster	0.23	0.17	0.52	0.05	0.05	0.02	good	0.18	−0.26	−0.22	0.01	0.04	0.01
disgust	−0.05	0.07	0.15	0.03	0.02	0.01	happiness	0.09	0.57	0.53	0.02	−0.01	0.01
evil	0.04	0.01	−0.14	0.05	0.02	0.01	heaven	0.01	−0.02	−0.11	0.05	0.04	0.03
failure	0.04	0.14	−0.12	0.07	0.06	0.05	joy	0.13	0.52	0.23	0.02	−0.02	−0.01
hate	−0.07	−0.18	0.06	0.02	−0.02	−0.01	laughter	0.18	0.23	0.26	0.05	0.02	0.03
horrible	−0.03	−0.05	0.45	0.05	0.03	0.01	love	0.10	0.43	0.21	0.02	−0.05	−0.03
nasty	0.01	0.45	0.60	0.02	0.01	0.01	marvelous	−0.01	−0.19	−0.32	0.07	0.08	0.08
pain	−0.14	−0.26	−0.35	0.06	0.06	0.03	peace	0.04	0.43	0.08	0.02	−0.01	−0.03
terrible	0.12	−0.05	0.21	0.04	0.03	0.03	pleasure	0.01	0.41	−0.01	0.01	0.01	−0.01
ugly	−0.09	−0.07	0.06	0.01	0.02	0.01	wonderful	0.12	−0.37	−0.56	0.04	0.08	0.06
M (SD)	−0.05 (0.14)	−0.13 (0.42)	−0.02 (0.47)	0.05 (0.03)	0.04 (0.04)	0.03 (0.04)		0.06 (0.10)	0.14 (0.36)	0.03 (0.30)	0.03 (0.02)	0.02 (0.04)	0.02 (0.04)
Dark chocolate							*Milk* chocolate						
Dark 1	−0.44	−0.41		−0.10	−0.11		Milk 1	−0.10		−0.30	−0.04		−0.07
Dark 2	0.10	−0.36		−0.10	−0.11		Milk 2	0.01		−0.28	−0.07		−0.08
Dark 3	−0.10	−0.14		−0.07	−0.08		Milk 3	−0.08		−0.31	−0.06		−0.07
Dark 4	−0.15	−0.23		−0.07	−0.10		Milk 4	−0.17		−0.38	−0.05		−0.10
Dark 5	−0.15	−0.41		−0.11	−0.10		Milk 5	0.16		−0.37	−0.06		−0.08
Dark 6	−0.13	−0.18		−0.09	−0.11		Milk 6	0.17		−0.38	−0.05		−0.10
Dark 7	−0.18	−0.27		−0.10	−0.11		Milk 7	−0.03		−0.22	−0.05		−0.08
M (SD)	−0.15 (0.16)	−0.29 (0.10)		−0.09 (0.02)	−0.10 (0.01)			−0.01 (0.13)		−0.32 (0.06)	−0.05 (0.01)		−0.08 (0.01)

A significant effect of the categories of the stimuli on the easiness estimates was found in the IAT ($F(4, 36) = 3.40$, $p = 0.02$), while in both the SC-IATs this effect was not significant (Dark SC-IAT: $F(3, 30) = 2.81$, $p = 0.06$ and Milk SC-IAT: $F(3, 30) = 1.98$, $p = 0.14$).

In the IAT, the target object *Dark* was the most difficult category ($B = -0.15$, $SE = 0.$ $t(36) = -3.05$, $p < .001$). No significant effects were found for ~~all the~~ other categories of stimuli ($B_{\text{Milk}} = -0.01$, $SE = 0.05$, $t(36) = -0.16$, $p = 0.89$, $B_{\text{Bad}} = -0.05$, $SE = 0.04$, $t(36) = -1.28$, $p = 0.20$, and $B_{\text{Good}} = 0.06$, $SE = 0.04$, $t(36) = -1.62$, $p = 0.11$).

Despite in both the SC-IATs the overall effect of the categories of the stimuli was not significant, a significant effect of the target object categories was found in both the Dark SC-IAT ($B_{\text{Dark}} = -0.29$, $SE = 0.13$, $t(30) = -2.15$, $p = 0.03$) and ~~in~~ the Milk SC-IAT ($B_{\text{Milk}} = -0.31$, $SE = 0.13$, $t(30) = -2.41$, $p = 0.02$). In both cases, the target objects tended to be the most difficult stimuli. The effect of the evaluative dimensions was significant in neither the Dark SC-IAT ($B_{\text{Bad}} = -0.13$, $SE = 0.10$, $t(30) = -1.35$, $p = 0.19$, and $B_{\text{Good}} = 0.14$, $SE = 0.10$, $t(30) = -1.40$, $p = 0.17$), nor ~~in~~ the Milk SC-IAT ($B_{\text{Bad}} = -0.02$, $SE = 0.10$, $t(30) = -0.17$, $p = 0.87$, and $B_{\text{Good}} = 0.03$, $SE = 0.10$, $t(30) = 0.36$, $p = 0.72$).

In all implicit measures, there were stimuli showing easiness estimates far away from the estimates of the stimuli belonging to the same category, although the pattern was not consistent between implicit measures. take for example the stimulus *glory* (category *Good*). In the IAT, it resulted to be a particularly difficult stimulus, also in respect to the average level of easiness of the stimuli belonging to the same category. In both SC-IATs, it resulted to be a particularly easy stimulus.

Log-normal models

Model comparison is reported in Table 7.2 (i.e., Model identified by capital "T"). Model T2 produced aberrant estimates (i.e., correlation between stimuli random slopes equal to one) in all implicit measures, suggesting a low within–stimuli between–conditions variability. Model T3 was the best fitting model for all implicit measures. Consequently, condition–specific respondents' speed estimates (τ_{DGMB}, τ_{MGDB}, τ_{DG}, τ_{DB}, τ_{MG},τ_{MB}) and overall stimuli estimates δ_{sm} of the log-normal model were obtained for each implicit measure.

Faster responses were observed in the MGDB condition ($B = -0.35$, $SE = 0.01$), in the DB condition ($B = -0.45$, $SE = 0.02$), and in the MG condition ($B = -0.48$, $SE = 0.01$), than in their respective contrasting conditions ($B_{\text{DGMB}} = -0.11$, $SE = 0.02$, $B_{\text{DG}} = -0.44$, $SE = 0.03$, and $B_{\text{MB}} = -0.43$, $SE = 0.01$, for the IAT, the Dark SC-IAT, and the Milk SC-IAT, respectively). Respondents showed similar variability in the two IAT conditions ($\sigma^2_{\text{DGMB}} = 0.05$ and $\sigma^2_{\text{MGDB}} = 0.03$), as well as similar variability in the two Milk SC-IAT conditions ($\sigma^2_{\text{MB}} = 0.02$ and $\sigma^2_{\text{MG}} = 0.01$). The variability in the two Dark SC-IAT conditions was the same ($\sigma^2 = 0.02$). Stimuli variability was extremely low for all three measures (0.004, 0.004, and 0.003 for the IAT, the Dark SC-IAT, and the Milk SC-IAT, respectively).

The estimates of the stimuli time intensity for the IAT, the Dark SC-IAT, and the Milk SC-IAT are reported in Table 7.3. A significant effect of the categories of the stimuli on the stimuli time intensity was found in all implicit measures (IAT: $F(4, 36) = 63.49$, $p < .001$, Dark SC-IAT: $F(3, 30) = 28.05$, $p < .001$, and Milk SC-IAT: $F(3, 30) = 20.57$, $p < .001$).

In the IAT, the target object *Dark* was the category of stimuli requiring less time for getting a response ($B = -0.09$, $SE = 0.01$, $t(36) = -11.10$, $p < .001$), immediately followed by the target object *Milk* ($B = -0.05$, $SE = 0.01$, $t(36) = -6.43$, $p < .001$). Both evaluative dimensions tended to require more time for getting a response ($B_{\text{Bad}} = 0.05$, $SE = 0.01$, $t(36) = 8.25$, $p < .001$, and $B_{\text{Good}} = 0.03$, $SE = 0.01$, $t(30) = 4.62$, $p < .001$).

In both the SC-IATs, significant effects were found for the corresponding target objects ($B_{\text{Dark}} = -0.10$, $SE = 0.01$, $t(30) = -8.05$, $p < .001$ and $B_{\text{Milk}} = -0.08$, $SE = 0.01$, $t(30) = -6.92$, $p < .001$) and on the evaluative dimension *Bad* (Dark SC-IAT: $B = 0.04$, $SE = 0.01$, $t(30) = 3.91$, $p < .001$ and Milk SC-IAT: $B = 0.03$, $SE = 0.01$, $t(30) = 3.21$, $p < .001$). The target objects required less time for getting a response, while the evaluative dimension *Bad* required more time. The effect of the evaluative dimension *Good* was significant in neither the Dark SC-IAT ($B = 0.02$, $SE = 0.01$, $t(30) = 1.99$, $p = .05$) nor ~~in~~ the Milk SC-IAT ($B = 0.02$, $SE = 0.01$, $t(30) = 1.88$, $p = .07$).

7.2.2 Comprehensive models

Rasch models

Models A1^{C}, A2^{C}, and A3^{C} were compared between each other. Model A3^{C} (AIC = 22,365, BIC = 22,618, Log-likelihood = −11,154, Deviance = 22,309) resulted as the best fitting one (Model A1^{C}: AIC = 22,991, BIC = 23,036, Log-likelihood = −11,490, Deviance = 22,981, Model A2^{C}: AIC = 22,906, BIC = 22,996, Log-likelihood = −11,223, Deviance = 22,886), providing condition–specific respondents' ability estimates for each implicit measure (θ^{C}_{DGMB}, θ^{C}_{MGDB}, θ^{C}_{DG}, θ^{C}_{DB}, θ^{C}_{MG}, and θ^{C}_{MB}), as well as overall stimuli easiness estimates across implicit measures (b^{C}_{s}).

The highest probability of a correct response was observed in the MGDB condition (*log-odds* = 4.05, *SE* = 0.13), followed by that in the DB condition (*log-odds* = 3.45, *SE* = 0.11) and that in the MG condition (*log-odds* = 3.41, *SE* = 0.10). The probability of a correct response in the DG condition and that of a correct response in the MB condition were similar (DG: *log-odds* = 3.23, *SE* = 0.10 and MB *log-odds* = 3.21, *SE* = 0.10). The lowest probability of a correct response was observed in the DGMB condition (*log-odds* = 2.92, *SE* = 0.09). The MGDB condition was the one showing the highest respondents' variability ($\sigma^2 = 1.05$), followed by that in the DG condition ($\sigma^2 = 0.84$), and that in the MB condition ($\sigma^2 = 0.75$). Respondents' variability in the DG condition and that in the MG condition were similar ($\sigma^2_{DG} = 0.63$ and $\sigma^2_{MG} = 0.64$). The DGMB condition showed the lowest variability ($\sigma^2 = 0.47$). Variability at the stimuli level was 0.11.

The estimates of the easiness parameters from Model A5 are reported in Table 7.4. A significant effect of the categories of the stimuli on their easiness estimates was found ($F(4, 36) = 2.83$, $p = 0.04$). The target object *Dark* was the most difficult category ($B = -0.24$, *SE* = 0.11, $t(36) = -2.28$, $p = 0.03$). The target object *Milk* was fairly difficult as well, although its effect was not significant ($B = -0.17$, *SE* = 0.11, $t(36) = -1.57$, $p = 0.13$). Neither the effect of the category *Bad* ($B = -0.03$, *SE* = 0.08, $t(36) = -0.49$, $p = 0.63$) nor that of the category *Good* ($B = 0.14$, *SE* = 0.08, $t(36) = 1.85$, $p = 0.07$) were significant, although stimuli of the latter category tended to be the easiest ones.

Table 7.4: Stimuli easiness estimates (b_s) and time intensity estimates (δ_s) for the Comprehensive model.

	b	δ		b	δ
Bad attributes			*Good* attributes		
agony	-0.89	0.10	beautiful	-0.13	0.01
annoying	-0.74	0.13	excellent	0.10	0.06
bad	-0.02	0.03	glory	0.26	0.04
disaster	0.47	0.05	good	-0.07	0.03
disgust	0.11	0.03	happiness	0.55	0.01
evil	0.01	0.04	heaven	-0.01	0.05
failure	0.07	0.08	joy	0.44	0.01
hate	-0.05	0.01	laughter	0.36	0.04
horrible	0.18	0.04	love	0.37	-0.02
nasty	0.47	0.02	marvelous	-0.17	0.09
pain	-0.27	0.06	peace	0.27	0.01
terrible	0.17	0.05	pleasure	0.20	0.01
ugly	0.01	0.03	wonderful	-0.30	0.08
M (SD)	-0.04 (0.40)	0.05 (0.03)		0.14 (0.26)	0.03 (0.03)
Dark Chocolate			*Milk* Chocolate		
Dark 1	-0.49	-0.10	Milk 1	-0.22	-0.04
Dark 2	-0.14	-0.10	Milk 2	-0.14	-0.07
Dark 3	-0.13	-0.07	Milk 3	-0.21	-0.06
Dark 4	-0.21	-0.08	Milk 4	-0.30	-0.07
Dark 5	-0.32	-0.10	Milk 5	-0.09	-0.06
Dark 6	-0.17	-0.10	Milk 6	-0.10	-0.07
Dark 7	-0.25	-0.10	Milk 7	-0.12	-0.06
M (SD)	-0.24 (0.13)	-0.09 (0.01)		-0.17 (0.08)	-0.06 (0.01)

Log-normal models

Models T1$^{\text{C}}$, T2$^{\text{C}}$, and T3$^{\text{C}}$ were compared between each other. Model T3$^{\text{C}}$ (AIC $=$ 38,933, BIC $=$ 39,195, Log-likelihood $=$ $-$19,437, Deviance $=$ 38,875) resulted as the best fitting one (Model T1$^{\text{C}}$: AIC $=$ 44,624, BIC $=$ 44,678, Log-likelihood $=$ -22,306, Deviance $=$

44,612, and Model $T2^C$: AIC = 43,448, BIC = 43,548, Log-likelihood = −22,306, Deviance = 44,612).

The responses in the IAT associative conditions tended to be slower ($B_{\text{DGMB}} = -0.12$, $SE = 0.02$ and $B_{\text{MGDB}} = -0.35$, $SE = 0.02$) than in the Dark SC-IAT associative conditions ($B_{\text{DB}} = -0.47$, $SE = 0.02$ and $B_{\text{DG}} = -0.45$, $SE = 0.02$) and in the Milk SC-IAT associative conditions ($B_{\text{MB}} = -0.45$, $SE = 0.02$ and $B_{\text{MG}} = -0.50$, $SE = 0.01$).

Respondents' variability was slightly higher in the IAT conditions ($\sigma^2_{\text{DGMB}} = 0.05$ and $\sigma^2_{\text{MGDB}} = 0.03$) than in both the associative conditions of the SC-IATs. Respondents showed the same variability in the Dark SC-IAT associative conditions ($\sigma^2 = 0.02$), and a slightly different variability in the Milk SC-IAT associative conditions ($\sigma^2_{\text{MB}} = 0.02$ and $\sigma^2_{\text{MG}} = 0.01$). Stimuli variability was extremely low ($\sigma^2 = 0.004$).

Time intensity estimates of Model $T3^C$ are reported in Table 7.4. A significant effect of the categories of the stimuli on their time intensity estimates was found ($F(4, 36) = 42.93$, $p < .001$). The target objects categories required the least amount of time for getting a response ($B_{\text{Dark}} = -0.09$, $SE = 0.01$, $t(36) = -8.79$, $p < .001$ and $B_{\text{Milk}} = -0.06$, $SE = 0.01$, $t(36) = -5.85$, $p < .001$) than both the evaluative dimensions ($B_{\text{Bad}} = 0.05$, $SE = 0.01$, $t(36) = 6.49$, $p < .001$ and $B_{\text{Good}} = -0.09$, $SE = 0.01$, $t(36) = 4.25$, $p < .001$).

7.2.3 Relationship between model estimates and typical scoring

Since condition–specific respondents' estimates were available for both the Rasch model and the log-normal model, differential measures for ability and speed estimates were computed. These measures express the bias on respondents' accuracy or speed performance due to the effect of the associative conditions. Ability differential measures were computed so that positive scores stood for a higher ability in the DGMB condition than in the MGDB condition, or a higher ability in the associative condition where the target chocolate was associated with positive exemplars in both SC-IATs (the DG condition and the MG condition). Speed differential measures were computed so that positive scores stood for higher speed in the DGMB condition than in the opposite one, or higher speed in the condition where the target

chocolate was associated with positive attributes rather than with negative attributes.

A stepwise approach with forward selection was followed. The differential measures and their respective single estimates components were entered in different models to avoid collinearity. The Null model against which all full models were compared included only the intercept (i.e., expected average of the typical score). The predictors included in the Full models for each implicit measure are summarized in Table 7.5, as well as the predictors resulting from stepwise forward selection.

Table 7.5: Relations between typical scoring and model estimates.

		Predictors	*B*	*SE*	*Adjusted* R^2	Predictors	*B*	*SE*	*Adjusted* R^2
					Estimates of Single measure models				
IAT		Full model	*D* score $\sim \theta_{DGMB} + \theta_{MGDB} + \tau_{DGMB} + \tau_{MGDB}$				*D* score $\sim (\theta_{DGMB} - \theta_{MGDB}) + (\tau_{MGDB} - \tau_{DGMB})$		
		Null - Intercept	-0.58^{***}	0.04	0.00				
		Intercept	0.07	0.10	0.89	Intercept	0.05^{*}	0.03	0.89
		θ_{MGDB}	-0.14^{***}	0.03		$(\tau_{MGDB} - \tau_{DGMB})$	2.02^{***}	0.08	
		θ_{DGMB}	0.16^{***}	0.04		$(\theta_{DGMB} - \theta_{MGDB})$	0.15^{***}	0.03	
		τ_{DGMB}	-1.94^{***}	0.09					
		τ_{MGDB}	2.16^{***}	0.10					
Dark IAT	SC-	Full model	*D-Dark* $\sim \theta_{DG} + \theta_{DB} + \tau_{DG} + \tau_{DB}$				*D-Dark* $\sim (\theta_{DG} - \theta_{DB}) + (\tau_{DB} - \tau_{DG})$		
		Null - Intercept	-0.05^{**}	0.02	0.00				
		Intercept	0.11	0.07	0.82	Intercept	0.03^{**}	0.01	0.82
		τ_{DB}	3.51^{***}	0.15		$(\tau_{DB} - \tau_{DG})$	3.46^{***}	0.15	
		τ_{DG}	-3.35^{***}	0.16		$(\theta_{DG} - \theta_{DB})$	0.18^{***}	0.02	
		τ_{DG}	0.18^{***}	0.02					
		τ_{DB}	-0.18^{***}	0.02					
Milk IAT	SC-	Full model	*D-Milk* $\sim \theta_{MG} + \theta_{MB} + \tau_{MG} + \tau_{MB}$				*D-Milk* $\sim (\theta_{MG} - \theta_{MB}) + (\tau_{MB} - \tau_{MG})$		
		Null - Intercept	0.32^{***}	0.03	0.00				
		Intercept	-0.31^{*}	0.16	0.30	Intercept	0.21^{***}	0.03	0.25
		τ_{MB}	1.66^{***}	0.27		$(\tau_{MB} - \tau_{MG})$	1.77^{***}	0.28	
		τ_{MG}	-2.23^{***}	0.31		$(\theta_{MG} - \theta_{MB})$	0.13^{***}	0.03	
		θ_{MG}	0.16^{*}	0.03					
		θ_{MB}	-0.09^{*}	0.02					
					Estimates of Comprehensive models				
IAT		Full model	*D* score $\sim \theta^{C}_{DGMB} + \theta^{C}_{MGDB} + \tau_{DGMB} + \tau^{C}_{MGDB}$				*D* score $\sim (\theta^{C}_{DGMB} - \theta^{C}_{MGDB}) + (\tau^{C}_{MGDB} - \tau^{C}_{DGMB})$		

Table 7.5: Relations between typical scoring and model estimates.

		Predictors	*B*	*SE*	*Adjusted* R^2	Predictors	*B*	*SE*	*Adjusted* R^2
		Intercept	0.05	0.10	0.88	Intercept	0.03	0.03	0.88
		τ^{C}_{MGDB}	2.18^{***}	0.10		$(\tau^{C}_{MGDB} - \tau^{C}_{DGMB})$	2.04^{***}	0.08	
		τ^{C}_{DGMB}	-1.95^{***}	0.09		$(\theta^{C}_{DGMB} - \theta^{C}_{MGDB})$	0.11^{***}	0.02	
		θ^{C}_{DGMB}	0.13^{***}	0.04					
		θ^{C}_{MGDB}	-0.12^{***}	0.02					
Dark IAT	SC-	Full model	$D\text{-}Dark \sim \theta^{C}_{DG} + \theta^{C}_{DB} + \tau^{C}_{DG} + \tau^{C}_{DB}$			$D\text{-}Dark \sim (\theta^{C}_{DG} - \theta^{C}_{DB}) + (\tau^{C}_{DB} - \tau_{DG})^{C}$			
		Intercept	0.04	0.08	0.78	Intercept	0.03^{**}	0.01	0.78
		τ^{C}_{DB}	3.52^{***}	0.18		$(\tau^{C}_{DB} - \tau_{DG})^{C})$	3.49^{***}	0.17	
		τ^{C}_{DG}	-3.40^{***}	0.19		$(\theta^{C}_{DG} - \theta^{C}_{DB})$	0.14^{***}	0.02	
		θ^{C}_{DG}	0.15^{***}	0.02					
		θ^{C}_{DB}	-0.14^{***}	0.02					
Milk IAT	SC-	Full model	$D\text{-}Milk \sim \theta^{C}_{MG} + \theta^{C}_{MB} + \tau^{C}_{MG} + \tau^{C}_{MB}$			$D\text{-}Milk \sim (\theta^{C}_{MG} - \theta^{C}_{MB}) + (\tau^{C}_{MB} - \tau^{C}_{MG})$			
		Intercept	-0.38^{*}	0.15	0.31	Intercept	0.21^{***}	0.02	0.25
		τ^{C}_{MB}	1.67^{***}	0.28		$(\tau^{C}_{MB} - \tau^{C}_{MG})$	1.82^{***}	0.29	
		τ^{C}_{MG}	-2.27^{***}	0.32		$(\theta^{C}_{MG} - \theta^{C}_{MB})$	0.12^{***}	0.03	
		θ^{C}_{MB}	-0.08^{*}	0.04					
		θ^{C}_{MG}	0.17^{***}	0.04					

Note: *** $p < .001$, ** $p < .01$. θ Ability estimates, τ: Speed estimates, DGMB: Dark/Good-Milk/Bad condition (IAT), MGDB: Milk/Good-Dark/Bad condition(IAT), DG: Dark/Good condition (Dark SC-IAT), DB: Dark/Bad condition (Dark SC-IAT), MG: Milk/Good condition (Milk SC-IAT), MB: Milk/Bad condition (Milk SC-IAT), C: Estimates obtained with the Comprehensive models.

The estimates of the intercepts of the Null models were significantly different from 0. The estimates of the intercepts of the *D-Milk* and that of the *D* score showed larger effect sizes than that of the *D-Dark* .

Forward selection always pointed the full models as the models best accounting for the typical scoring. Ability estimates were always retained in the models. However, the effect size of their coefficients was smaller than that of the coefficients of the speed estimates.

The linear combination of estimates of the Single measure models and their differential measures explained the same amount of variance of both the *D* score and the *D-Dark* . Differential measures explained a lesser proportion of variance of the *D-Milk* than that explain by the linear combination of their single components. Additionally, the *D-Milk* showed the smallest proportion of explained variance, regardless of the type of predictors.

Similar results were obtained for the estimates of the Comprehensive model. The proportion of variance of the *D-Dark* explained by Comprehensive model estimates was slightly lower that that explained by the estimates of the Single measure model. This result held for both the linear combination of the condition–specific estimates and their differential measures. Concerning the *D* score and the *D* Milk, the proportion of explained variance was almost identical to that explained by the estimates of the Single measure model.

7.2.4 Prediction of a behavioral outcome

The predictive ability of the Rasch model and the log-normal model estimates and that of the typical scoring were investigated and compared. Since condition–specific ability estimates and condition–specific speed estimates were available for both the Single measure models and the Comprehensive models, the predictive ability of the differential measures was investigated as well.

Results of the stepwise logistic regressions are reported in Table 7.6. Forward selection retained only the IAT *D* score, regardless of whether it was paired with the scores of each SC-IAT or with the *D-Sciat*. Also when considering the single components of the typical scoring methods, only the single components of the IAT were retained in the model. This

Table 7.6: Stepwise forward selection results: Choice prediction.

Model		*log-odds*	*SE*	*Nagelkerke* R^2	*Gen*	*DCC*	*MCC*
Null	Intercept	-0.08	0.16	0.00	0.48	0.00	1.00
		Typical scoring					
1	Intercept	-1.03^{***}	0.31	0.16	0.64	0.62	0.67
	D score	-1.55^{***}	0.39				
2	Intercept	-0.36	0.91	0.12	0.62	0.67	0.56
	M_{DGMB}	0.01^{**}	0.01				
	M_{MGDB}	-0.01^{*}	0.01				
		Single measure models					
3	Intercept	-0.97^{***}	0.29	0.16	0.64	0.68	0.60
	$(\tau_{\text{MGDB}} - \tau_{\text{DGMB}})$	-3.66^{***}	0.93				
4	Intercept	0.14	0.73	0.19	0.64	0.66	0.63
	τ_{DGMB}	2.83^{**}	1.04				
	τ_{MGDB}	-5.77^{***}	1.66				
	τ_{DG}	4.49^{*}	2.23				
		Comprehensive models					
5	Intercept	-0.95^{***}	0.29	0.15	0.64	0.67	0.60
	$(\tau^{\text{C}}_{\text{MGDB}} - \tau^{\text{C}}_{\text{DGMB}})$	-3.57^{***}	0.91				
6	Intercept	0.44	0.81	0.19	0.64	0.65	0.63
	$\tau^{\text{C}}_{\text{DGMB}}$	2.45^{*}	1.08				
	$\tau^{\text{C}}_{\text{MGDB}}$	-5.99^{***}	1.76				
	$\tau^{\text{C}}_{\text{DG}}$	5.29^{*}	2.59				

Note: $^{***}\ p < .001$, $^{**}\ p < .01$, $^{*}\ p < .05$. θ: Ability estimates, τ: Speed estimates, DGMB: Dark/Good-Milk/Bad condition (IAT), MGDB: Milk/Good-Dark/Bad condition (IAT), DG: Dark/Good condition (Dark SC-IAT), DB: Dark/Bad condition (Dark SC-IAT), MG: Milk/-Good condition (Milk SC-IAT), MB: Milk/Bad condition (Milk SC-IAT), *Gen*: General accuracy, *DCC*: Dark chocolate choice accuracy, *MCC*: Milk chocolate choice accuracy, C: Estimates obtained with a comprehensive modeling of IAT and SC-IAT responses.

model explained the lowest proportion of variance, and it resulted in the lowest MCC of all models.

The differential measures obtained from the difference between the condition–specific speed estimates of the IAT were retained by forward selection, both for the Single measure models and for the Comprehensive model. These models explained a slightly lower proportion of variance than that explained by the models including their linear components counterparts.

For both the Single measure models and the Comprehensive model, forward selection retained the speed estimates of the IAT associative conditions (i.e., τ_{DGMB} and τ_{MGDB}) and that of DG associative condition (i.e., τ_{DG}). These models explained a higher proportion of variance than that explained by the *D* score and its linear components.

With the only exception of the model including the single components of the IAT *D* score, all other models showed the same General accuracy of prediction.

7.3 Final remarks

The approach presented in this study represents a first attempt at a comprehensive modeling of the IAT and the SC-IAT. The Rasch model and the log-normal model estimates resulting from the application of the (G)LMMs provided interesting insights on the functioning of these implicit measures, and on the consequences of not accounting for the non-independence of the observations.

The stimuli estimates obtained with both the Single measure models and the Comprehensive models allowed for identifying stimuli with estimates far away from the estimates of the stimuli belonging to the same category. In the Single measure models, measure–specific stimuli estimates allowed for highlighting stimuli with a different functioning in all implicit measures. However, the pattern of these stimuli was not consistent between implicit measures. For example, stimulus *good* was a particularly easy stimulus in the IAT as well as a particularly demanding one in both SC-IATs. The stimuli estimates obtained with the Comprehensive models were less extreme than the ones obtained with the Single measure models. The estimates obtained with the latter approach might be artificially inflated by unaccounted and uncontrolled sources of error variance. By controlling for the sources of error variance

between implicit measures, the stimuli estimates of the comprehensive modeling approach might provide more reliable stimuli estimates, describing the stimuli functioning across implicit measures.

Having measure–specific stimuli is useful nonetheless, and allows for investigating the stimuli functioning according to the specific measure in which they are administered. However, the Single measure approach risks for resulting in biased stimuli estimates, that might end in misleading inferences on the stimuli functioning. Potentially, measure–specific stimuli estimates can be obtained by specifying stimuli random slopes in each implicit measure. Since also respondents' variability is of interest, their random slopes in the associative conditions should be specified. Besides not being identified in a Rasch modeling framework (i.e., either respondents or stimuli have to be centered to 0), it is highly unlikely that a model of this complexity would converge because it would require a high within–stimuli between–measures variability.

Regardless of whether they were obtained with the Single measure models or the Comprehensive model, the ability estimates provided a lower contribution to the prediction of the typical scoring methods than speed estimates. Considering that the typical scoring methods are mostly based on time responses, this result is not surprising. Nonetheless, since each incorrect response is replaced with an inflated response time, also the accuracy performance of the respondents play a role in determining the final score.

The IAT *D* score was the classic score with the best predictive ability of the behavioral outcome. This result is not surprising given that the data are the same as those in Chapter 2. When the linear components of the typical scoring procedures were used to predict the choice, only the average response times of the two associative conditions of the IAT were retained in the model. This model resulted to be the one with the lowest predictive ability in respect to the Milk chocolate choice. Grounding on the results obtained with typical scoring procedures, it appears that only the IAT provides the information needed for predicting the choice.

The results obtained with the estimates of the Single measure models and those obtained with the Comprehensive models move in another direction. Regarding differential measures,

forward selection only retained the differential measure computed with the condition–specific speed estimates of the IAT. This result held for estimates obtained with the Single measure models and the Comprehensive models. However, when the linear combination of their single components was used to predict the choice, the contribution of the speed in the Dark-Good condition of the Dark SC-IAT was highlighted. Consequently, it can be speculated that the behavioral choice is driven more by the liking for Dark chocolate than by the dislike for Milk chocolate. By only considering the typical scoring methods, that are affected by different sources of error variance, or the differential measures, which are known to confound the contribution of their components (Fiedler, Messner, & Bluemke, 2006), it was not possible to disentagle the automatic association mostly involved in the prediction of the behavior.

Although the prediction provided by the Single measure model estimates and that provided by the Comprehensive model estimates do not result in higher accuracy of the choice, they do explain an higher proportion of variance of the choice. Most importantly, they allow for a deeper understanding of the processes underlying the actual choice.

In this study, implicit measures were used for the assessment of a quite trivial preference, namely the chocolate preference. It would be interesting to investigate whether this approach would replicate on implicit measures for the assessment of other preferences, like the Coke vs Pepsi one in Karpinski and Steinman (2006), but, most importantly, of socially relevant constructs, such as implicit prejudice. Given that the model estimates provide a deeper and more thorough understating of the processes underlying people's behaviors, this modeling framework might be used for shedding a new light on inter group behaviors such as the decision to affiliate with people belonging to socially stigmatized out-groups.

Finally, the validity of this approach could be directly tested by designing implicit measures with *a priori* malfunctioning stimuli. If the approach is valid, it should be able to pinpoint the malfunctioning stimuli.

Chapter 8

Conclusions

This thesis was aimed at finding new methods for more rigorous analyses of implicit measures data by following three paths. The sound path composes the main part of the thesis. It was aimed at finding measurement models for analysis of the IAT and the SC-IAT, both when they are administered as stand-alone measures and when they are administered concurrently. The fair path took a direction consistent with the typical approach to the analysis of implicit measures data. It was aimed at introducing new scoring algorithms able to align the differences in the scoring procedures of the IAT and the SC-IAT. The alignment of the scoring differences allows for a fairer comparison between the predictive performance of the two implicit measures. Finally, the easy path aimed at improving the replicability of implicit measures results by providing new, open source tools for computing the IAT and the SC-IAT scores.

In this chapter, the main findings, implications, and limitations of the sound path and fair path are discussed. A comment on the overall ability of the thesis to meet the final aim closes the argumentation.

8.1 The sound path

The first step of the sound path was to find an appropriate modeling framework for the analysis of the IAT data. The measure obtained from the IAT strongly depends on the functioning

of the stimuli used to represent the categories (e.g., Bluemke & Friese, 2006). Consequently, a modeling approach resulting in stimulus–specific information appeared to be the most appropriate modeling approach for gaining a better understanding on the functioning of the IAT. Specifically, we were looking for a model able to disentangle the contribution of respondents' characteristics from that of the task in determining the observed responses. While reviewing the modeling frameworks proposed for the analysis of the IAT data, the lack of an approach able to provide such a fine-grained information at the stimuli level was blatant. The approaches introduced so far for the analysis of the IAT data do provide extremely useful information on the cognitive processes that underlie the performance at the IAT. They all point at the same direction: The IAT effect cannot be considered as just the expression of implicit processes, but it also includes components dependent on controlled processes that have to be taken into account for drawing meaningful conclusions from IAT data. Particular caution should be paid when the typical *D* score is used for scoring the IAT. The *D* score confounds the contribution of automatically activated associations with that of other processes, such as the effort to overcome automatically activated bias (Conrey et al., 2005) or to use other strategies to simplify the task (Klauer et al., 2007; Meissner & Rothermund, 2013).

However, none of these modeling frameworks was able to provide the information at the stimuli level we were looking for, nor they could disentangle the contribution of the task from that of the respondents in determining the observed responses. Concerning the stimuli, the most fine-grained information provided by these models is at the stimuli categories level. Some of these models, like the Diffusion Model or the Discrimination-Association model, provided parameters that were a mixture of task difficulty and respondents' ability, in sharp contrast with the peculiarities we were looking for.

Given that the aim was to disentangle the respondent's component from the task component, and, specifically, to gain information at the levels of the individual stimulus and the individual respondent, a Rasch framework for the analysis of the IAT data represented the best modeling approach. Evidence from previous study already showed the effectiveness of the application of the Many Facet Rasch Model (MFRM) to IAT data for providing a fine-grained information at the level of the individual stimulus. However, also this solution presented some

drawbacks that could not be ignored. Firstly, the MFRM was applied to the discretized response times of the IAT. The discretazion of a continuous variable results in a potentially large loss of information. Besides, the decision on the number of quantiles into which the continuous variable should be discretized plays an important role and might influence the results. Secondly, the fully-crossed structure of the IAT was overlooked. The fully-crossed design characterizing the IAT and the SC-IAT, and all experiments in which the same set of stimuli is presented multiple times to the same sample of respondents in different conditions, produces sources of random variability at the level of the single observations. The sources of random variability generate dependencies between the single observations which in turn break the assumption of local independence on which the Rasch model and the log-normal model are based. The MFRM can address other sources of variability than just the ones due to respondents' ability and stimuli difficulty, such as the variability due to the associative conditions of the IAT. However, there are reasons to believe that the sources of variability and related dependencies go beyond the respondents, the stimuli, and the associative conditions.

Despite the shortcomings highlighted for the application of the MFRM, a Rasch approach to IAT data represented the choice most in line with the aim of the sound path. Some pieces of the puzzle were still missing nonetheless. The issue of the sources of variability in the data remained, and the need for a methodology able to address them was urgent. Moreover, the MFRM was applied only on the discretized response times, hence the information retrievable from accuracy responses was disregarded. A modeling framework able to consider both accuracy and time responses, even in separate models, would allow for potentially gathering all the information from the IAT data. Finally, a modeling framework flexible enough to include other implicit measures administered together with the IAT, or as stand-alone measures, is a step forward to the modeling of implicit measures.

Summarizing, we were looking for a modeling framework able to provide a Rasch parametrization of both accuracy and time responses considered in their continuous nature, to account for the sources of variability at the level of the single observations, and with the possibility of being extended to model multiple implicit measures at the same time.

Linear Mixed-Effects Models (LMMs) are the modeling framework that meets all the

above mentioned requirements. Their ability of addressing the sources of dependencies in the data and their flexibility for being extended to multiple measures are their most outstanding an obvious features. A less obvious and less straightforward feature of LMMs is their link with the Rasch model, and specifically, how LMMs allows for estimating the parameters of this Psychometrics model. However, it must be considered that the Rasch model is nothing else than a linear model for latent trait variables. The link between Generalized LMMs (GLMMs) and the Rasch model becomes blatant when the equation of the Rasch model and that of the inverse link function of a Generalized Linear Model (GLM, $logit^{-1}$) are compared. The only difference concerns the interpretation of the parameters, and the relationship between the characteristics of the respondents and those of the stimuli. While in original formulation of the Rasch model they move in opposite directions, so that the stimulus could be considered as a sort of impediment (i.e., difficulty) for the response, in the application of the GLM, they move in the same direction. Consequently, the stimulus parameter can be considered as a facilitation property of the stimulus (i.e., easiness).

By including the matrix that defines the random effects into the linear component of the model, the structure of the GLM can be extended to be a GLMM. GLMMs allow for obtaining a Rasch parametrization of the data while acknowledging the fully-crossed structure of the IAT and its related sources of dependency.

Nonetheless, by applying GLMMs to accuracy responses, only a Rasch parametrization of the accuracy responses is obtained. Accuracy responses contain just a part of information, while time responses are expected to convey the highest amount of information. Considering the normal density distribution of the log-transformed time responses allows for avoiding the discretization needed for the application of the MFRM and results in the estimation of the log-normal model parameters (van der Linden, 2006). The log-normal model is a model for response times which yields a parametrization of the data similar to that provided by the Rasch model. Specifically, the observed responses can be explained by considering a respondent characteristic (i.e., speed parameter) and a stimulus characteristic (i.e., time intensity parameter). The log-normal model estimates can be obtained by applying LMMs to the log-time responses of the IAT.

The parameters of the Rasch and log-normal models are obtained from the random structures defined in each (G)LMMs. Different random structures yield different parametrization of the data, according to the random factor on which the multidimensionality is allowed on, either the respondents or the stimuli. Models with different random structures have been specified for the analysis of the accuracy and log-time responses of the IAT. The feasibility of these models, their usefulness for the analysis of IAT data, and their comparison with typical IAT scoring methods were tested in two studies employing two different IATs (see Chapter 5).

In a first study, a Race IAT was used. Regarding accuracy responses, the best fitting model was the one where the multidimensionality was allowed at the stimuli level, while respondents were centered at 0. The random structure of this model yielded condition–specific stimuli estimates and overall across–conditions respondents estimates. Consequently, condition–specific easiness stimuli estimates and overall respondents ability estimates of the Rasch model were obtained. The condition–specific estimates of the stimuli can be used for investigating the contribution of each stimulus to the IAT effect. The fact that the best fitting model was the one allowing for the multidimensionalty at the stimuli level means that there was a high within–stimuli between–conditions variability, along with a low within–respondents between–conditions variability. The functioning of the stimuli did change according to the associative conditions in which they were presented. It implies that the functioning of the stimuli changed according to the category of stimuli with which they shared the response key. In this instance, all stimuli tended to be easier in the White-Good/Black-Bad condition than in the opposite one. *Good* evaluative attributes were the stimuli showing the highest difference between the two conditions, immediately followed by *Bad* evaluative attributes. The stimuli representing Black people faces were the stimuli giving the least contribution to the IAT effect. Drawing on these results, the IAT effect appeared to be mostly driven by the evaluative dimensions, specifically by the positive one. These results are in line with those found with previous applications of the MFRM to the IAT data, according to which the IAT effect is mostly driven by positive attributes, and, as such, it should be interpreted as the expression of ingroup preference rather than outgroup derogation (positive primacy effect; e.g.,

Anselmi et al., 2013).

This result is further corroborated by the low difference between the condition–specific estimates of the category *Black*. As such, the easiness of categorization of these stimuli did not change much depending on the evaluative dimension with which they shared the response key. It can be speculated that Black people were neither strongly associated with negative attributes nor with positive ones, and that the resulting IAT effect was mostly driven by the evaluations made on White people faces.

The best fitting model for the log-time responses was the one allowing for the multidimensionality at the level of the respondents, while stimuli were centered at 0. This model resulted in the estimation of condition–specific respondents' speed parameters and overall stimuli time intensity estimates. The best fitting model indicated that there was a high within–respondents between–conditions variability along with a low within–stimuli between–conditions variability. This implies that respondents' performance changed between the two associative conditions, while the functioning of the stimuli remained the same between conditions. In other words, the time each stimulus required for getting a response did not change according to the stimuli category with which they shared the response key. The overall time intensity estimates can inform about the within–stimuli categories variability, and hence about stimuli hetereogeneity. Specifically, stimuli displaying a time intensity estimate too far away from the time intensity estimates of the other stimuli belonging to the same category should be replaced to reduce both the within–categories variability and the between–stimuli variability.

The condition–specific respondents' speed estimates allowed for delving deeper on the association(s) driving the IAT effect. Additionally, they provided a differential measure similar to the D score that expresses the bias on the speed performance due to the effect of the associative conditions. This differential measure can be used for further analysis, such as the prediction of behavioral outcomes.

The first study brought evidence in favor of the usefulness and feasibility of the proposed modeling framework for the analysis of the IAT data. However, neither the usefulness of the information at the stimuli level nor that at the respondents' level were tested. If the stimuli estimates provided by these models inform on the stimuli giving the highest contribution to

the IAT effect, it should be possible to isolate and select them for obtaining better performing IATs. Indeed, selecting only the stimuli giving the highest contribution to the IAT effect would reduce the stimuli heterogeneity and consequently the across-trials variability. If this is true, even the *D* score would result in a more reliable measure of the implicit construct under investigation. Moreover, the Rasch and log-normal model estimates obtained from the application of the (G)LMMs to IAT responses are less affected by sources of error variance than the *D* score is. Consequently, they result in a better inference of the construct under investigation and, as such, they potentially lead to a better prediction of a behavioral outcome.

In the second study, the two above-mentioned points were directly tested by using an IAT for the assessment of the implicit preference for Dark or Milk chocolate (Chocolate IAT). Both the accuracy and the log-time models were replicated on this data set. Condition–specific stimuli easiness estimates and overall ability estimates of the Rasch model, and condition–specific respondents' speed estimates and overall stimuli time estimates of the log-normal model were obtained. Also in this case, all stimuli tended to be easier in one condition over the other. Specifically, stimuli tended to be easier in the Milk-Good/Dark-Bad condition, and *Good* evaluative attributes were the stimuli having the greatest impact on the IAT effect. Given this pattern of results, it is possible to speculate that it was more the liking for Milk chocolate than the dislike for Dark chocolate that drove the performance at the IAT.

By using dark and milk chocolate as target objects of the IAT, it was possible to reward the participation of the respondents with a free bar of dark or milk chocolate. Obviously, the free bar of chocolate was not just a reward for the respondents but also the behavioral task of the experiment. The choice was registered by the experimenter, and it was used for investigating the predictive ability of the model estimates and that of the *D* score. Specifically, the choice was predicted by both the differential measures and the linear combination of ~~the~~ their single components in different logistic regressions. Backward deletion and the accuracy of the choice prediction provided by the models were used to determine the predictors best accounting for the observed chocolate choice. The log-normal speed estimates outperformed the *D* score in the prediction of the behavioral outcome. The lower predictive ability of the *D* score was observed both in comparison to its own linear components (although they explained

a lower proportion of variance) and in respect to both the linear combination of the condition–specific speed estimates and their *speed-differential*. The *D* score and its linear components do include uncontrolled sources of error variance due to the multiple sources of random variability in the IAT data. On the other hand, these sources of variability are accounted for in the speed estimates. Since error variance is most unlikely related to behaviors (Meissner et al., 2019), it should not be surprising to find a lower predictive ability of the *D* score. The *speed-differential* resulted in a slightly lower predictive ability than that provided by the linear combination of its single components.

The second study also investigated the ability of the condition–specific easiness estimates to pinpoint the stimuli giving the highest (lowest) contribution to the IAT effect. Since across-trial variability due to the stimuli heterogeneity is one of the factors that mostly affects the computation of the IAT *D* score, the reduction of the stimuli heterogeneity by selecting a specific pool of stimuli should provide more reliable *D* scores. By selecting only the stimuli providing the highest contribution to the IAT effect or the ones providing the least contribution to the IAT effect, the number of trials was reduced to $1/3$ of the original starting trials pool. Two additional *D* scores were computed, one on the data set including the stimuli giving the highest contribution to the IAT effect, and one on the data set including the least informative stimuli. The *D* score computed on the high informative stimuli data set did show a slightly better performance than the *D* score computed on the low informative stimuli. This result brings further evidence on the sensitivity of the *D* score to the across-trial variability due to the heterogeneity of the stimuli, at the point that it does not even matter whether the highest informative stimuli or the lowest informative ones are selected, as long as the variability is reduced.

The (G)LMMs approach showed its feasibility and appropriateness also for modeling SC-IAT data within a Rasch framework. However, this result has to be contextualized into the specific context of this thesis, where both the IAT and the SC-IATs were administered together. By separately analyzing the data from the three implicit measures, there are still sources of error variance that are left free to bias the estimates of the parameters. However, if the SC-IAT is administered as a stand-alone measure, a Rasch parametrization of its accuracy

and log-time responses can be easily obtained with the modeling framework presented in this work.

The comprehensive modeling approach of the IAT and the SC-IAT highlighted an IAT effect on both accuracy and speed performance of the respondents in all implicit measures. This result might indicate that, despite respondents slowed down in one condition, their accuracy performance is still impaired in that condition. Consequently, it is not surprising to find ability estimates of both the Single measure models and the Comprehensive models in predicting the typical score of each implicit measure. Indeed, typical scoring of the IAT and the SC-IAT are based on both time and accuracy responses (i.e., error responses are replaced with the average response time added with a penalty). The higher the number of incorrect responses, the higher the number of trials whose response time is replaced with the inflated one, and, consequently, the higher the average response time. It logically follows that, if in one condition both the speed performance and the accuracy performance are impaired, the average response time will be higher due to the combined effect of the slower response times and the inflated error response times, resulting in a higher difference between the associative conditions and in a larger effect size. Nonetheless, the ability estimates had a smaller effect size in the prediction of the typical scoring than the speed estimates.

The difference in respondents' performance between the associative conditions due to their slowing down and/or to a higher number of mistakes might be ascribable to just a small set of stimuli. In this case, the difference is not entirely related to automatic evaluative associations but also to the peculiarities of the task. Therefore, understanding how and why the estimates of some stimuli are far away from the those of the stimuli belonging to the same category becomes of particular relevance for getting a better and deeper understanding of the measure obtained, and of the inferences that can be reasonably done. If a stimulus is correctly responded but requires a longer time, it influences the average response time, hence skewing the result. If a stimulus is incorrectly responded, its response time is replaced by the average response time in that condition added with a penalty. Either way, the effect size of the *D* score will be artificially inflated by the response time of just some of the stimuli, and the inferences based on that should be taken with caution. Nonetheless, in considering the results on the

stimuli functioning, it was not possible to rule out the effect of the associative conditions.

The typical scores of both SC-ATs were always cut out in the prediction of the behavioral outcome. This result held for both the typical differential scores and the linear combination of their single components. If one was called to draw conclusions on the contribution of the SC-IATs to the prediction of behavioral outcomes, he/she would have probably inferred that the SC-IATs do not give any contribution to the choice prediction. Consequently, only the measure obtained from the IAT would have been considered as relevant for predicting behaviors.

Indeed, also the differential measures obtained from the model parameters estimates, both with the Single measures models and the Comprehensive models, pointed in the same direction as the typical scoring methods. Only the differential measures obtained from the IAT condition–specific speed estimates have been found to predict the choice, while the differential measures obtained from the estimates of the SC-IAT did not contribute in predicting the choice. However, when the linear combination of the condition–specific speed estimates of each implicit measure was used for predicting the choice, the speed in the Dark-Good condition of the Dark SC-IAT entered and remained in the model.

These results point at the risks of using the *D* score as a measure of the implicit construct under investigation. As already discussed, the *D* scores, as well as their linear components, are affected by sources of uncontrolled error variance resulting from the data structure itself. The administration of multiple measures to the same respondents generates further sources of variability and dependencies. Consequently, these scores result in biased estimates which, in this specific case, are not able to highlight the contribution of each implicit measure in the prediction of a behavioral outcome. The conclusions drawn from such scores should hence be taken with cautions.

Moreover, the results obtained on the choice prediction from both the study in Chapter 7 and that in Chapter 5 highlighted the issue related to the use of differential measures. In both cases, regardless of the implicit measure under consideration or the modeling framework used, differential measure are less accurate in predicting the behavioral outcome than the linear combination of their respective single components. This result is more evident for

the speed estimates of the log-normal model. The differences between the typical scores of implicit measures and their linear components is less evident, probably because the prediction is already affected by other sources of error variance. In Chapter 5, the model including the single components of the *speed-differential* was the one resulting in the highest accuracy of prediction of the Milk chocolate choice. Milk Chocolate Choice was disregarded by the *D* score, its linear components, and the *speed-differential*. In Chapter 7, the model including the linear combination of the condition–specific speed estimates of each implicit measure was the only one able to highlight the contribution of the speed of the Dark-Good condition of the Dark SC-IAT. This model did not result in a higher predictive accuracy than the others, but it did explain a higher proportion of variance of the choice. Besides, it made possible to gain a better understanding of the processes underlying the choice.

In the former case, differential measures did not provide a good prediction of one of the possible outcomes. In the latter one, differential measures were not able to identify the contribution of the SC-IAT in predicting the choice. The speed in only one of the conditions of the SC-IAT was found to contribute to choice prediction. The differential measure computed between the speed of the two conditions of the Dark SC-IAT might have confounded their importance and relevance for choice prediction, pointing at a null contribution of the Dark SC-IAT. Remarkably, the contribution of the Dark SC-IAT was completely lost when the linear components of the typical scoring were used.

The lack of predictive ability of differential measures might be due to the nature of differential measures themselves, which is confounding the contribution of each single component used for the computation of the single score. The computation of differential measures results in reliable scores only when the two quantities used for the computation have the same weight in determining the final score. This can be true only if a series of assumption are met, and, in the IAT case, this rarely happens (Fiedler et al., 2006). Firstly, the two target categories are assumed to give the same exact contribution to the IAT effect. In other words, the liking for one of the target categories has to be as strong as the dislike for the opposite category. This logically implies that the zero point stands for the absence of any positive or negative attitudes toward both target objects. Secondly, also the evaluative dimensions and the target

objects are assumed to have the same impact on the IAT effect. This assumption is in line with the idea of treating the stimuli as a fixed factor, which implies that they all have the same impact on the observed scores. However, as extensively discussed in the first chapter, considering stimuli as fixed factors in the IAT case is a stretch, and the distinct contribution of each stimulus to the IAT effect should be taken into account. Finally, systematic and unsystematic sources of variability are assumed to affect respondents' performance across the two conditions in the same way.

The information on respondents' performance and stimuli functioning provided by the modeling framework proposed in this thesis can be used for verifying the assumptions that have to be met for the computation of reliable and meaningful differential measures. Results on the stimuli functioning clearly suggest that each stimulus does give a different contribution to the IAT effect, and that their variability differently affect the final score. For instance, all studies highlighted a higher time intensity estimate for the attribute stimuli than for the image stimuli. The former ones required less time for getting a response than the latter ones. Additionally, in some cases image stimuli tended to be easier than attribute stimuli. These results clearly point at a different processing of the stimuli according to their type (attributes or images), and, blatantly, they cannot have the same effect on the observed responses.

Moreover, the contribution of the stimuli to the IAT effect and the relationship between the respondents' condition–specific speed estimates and the *D* score suggest that the differential score is mostly driven by the performance in one of the two associative conditions. Consequently, it seems bold to assume that the liking for one of the target categories is as strong as the dislike for the other one. Especially in the IAT case, which rests its measure on the juxtaposition between two objects, there might be cases in which the preference (dislike) for one object is extremely strong, while the contrasting object is not related to any particular positive or negative evaluation. Therefore, it can be assumed neither that attitudes towards the two contrasting objects have the same importance for the final score, nor that stimuli are processed in the same way and have the same impact on the final differential measure.

Regarding the violation of the last assumption, the one regarding the sources of systematic variability affecting the two conditions in the same way, the description of the fully-crossed

structure of implicit measures provided in the first chapter should have already clarified why this assumption is not meant to hold. Moreover, also the conceptualization of the ReAL model presented in Chapter 3, according to which different controlled processes can differently affect respondents' performance in the two associative conditions, makes hard to believe that this assumption could hold. Finally, the results on respondents' ability and speed performance obtained from the Rasch and log-normal modeling of the implicit measures do point to a difference in respondents' variability between the conditions. As a consequence of the violation of these assumptions, differential measures might not represent the best choice for expressing the implicit psychological construct assessed by implicit measures.

Given that the results on the relationship between model estimates and typical scoring methods and those on the choice prediction are almost identical for the estimates obtained with the Single measure models and those obtained with the Comprehensive model, one might be wondering about the advantages of using the latter approach over the former one. The former approach does result in measure–specific stimuli estimates, which inform about the functioning of the stimuli in each implicit measure. Conversely, the Comprehensive model results in overall stimuli estimates across implicit measures, hence providing a general information of stimuli functioning across measures. However, the apparent advantage of the Single measure model of providing measure–specific stimuli estimates is also its major shortcoming, as already discussed. By not addressing the between–measures variability, the new sources of error variance related to the administration of multiple implicit measures are left free to bias the estimates. Moreover, since the estimates are obtained from separate and independent models, they cannot be compared between each other. A comparison between the respondents' performance in the implicit measures is meaningless if not dangerous in terms of inferences that can be made.

8.2 The fair path

The fair path appears to be in clear antithesis with what has been said so far about typical scoring of implicit measures. However, effect size measures are still the most common ways for

scoring implicit measures data, both when administered as stand-alone measures and when they are administered together. The resulting scores are then used for further analyses and/or for comparing the performance of the different implicit measures in respect to some criteria (e.g., prediction of behavioral outcomes). However, the differences in both the administration and the scoring procedures of implicit measures such as the IAT and the SC-IAT might directly affect the score obtained at each of them. If these scores are then used for comparing the IAT and SC-IAT performance on different criteria, the comparison might result affected by artifacts which are not directly related to the goodness of the implicit measures itself but to elements of minor importance. How one can be sure that the lesser predictive ability in respect to a behavioral outcome provided by the SC-IAT is truly ascribable to the measure itself and not to some minor features? By providing easy-to-compute and easy-to-interpret effect size measures with which typical users of these implicit measures are more familiar ~~with~~, the approach presented in the fair path might help in answering this question or, at the very least, in fostering a fairer comparison between the IAT and the SC-IAT. Summarizing, the fair path was aimed at providing rigorous and comparable scoring methods for different implicit measures without moving apart from the typical approach.

While it is true that different implicit measures do have features that make them unique, there are features that can be aligned in both their administration and their scoring. The alignment of these differences allows for a fairer comparison between the performance of implicit measures. This leads to mainly two advantages. Firstly, the performance of the respondents on different measures can be reasonably compared, and secondly the results on the comparison between implicit measures performance in respect to different criteria can be mostly ascribed to the measure and not to other artifacts.

The new scoring methods that have been implemented do not necessarily result in a higher accuracy of the prediction. They do point to a higher predictive ability of the IAT in respect to that of the SC-IAT. In this case, the better performance of the IAT can be more easily pinned to the measure itself and not to artifacts due to the differences of scoring and administration procedures. Moreover, by taking out the role of the scoring in potentially influencing the results, it is possible to make more accurate speculations on the reasons why the IAT does

show a better performance than the SC-IAT. In the study reported in Chapter 2, the higher predictive of the IAT in respect to the SC-IAT might be due to the dichotomous nature of the choice, which is more in line with the comparative measure provided by the IAT than the absolute one provided by the SC-IAT.

Limitations and future directions

The modeling framework introduced in this thesis provides interesting and useful information on the functioning of different implicit measures, concerning both the respondents and the stimuli. For instance, it was possible to pinpoint the stimuli that gave the highest contribution to the IAT effect. This information can be further used for getting a better understanding of the automatic association(s) implicated in the performance at the IAT. Moreover, the information at the stimuli level help in reducing the across-trial variability by selecting only the most informative stimuli. By doing so, the administration time of the IAT can be reduced. At the respondents' level, it was possible to shed a new light on the components included into the *D* score, and to obtain a better inference on the implicit constructs under investigation.

However, the information yielded from the accuracy responses completely ignores the information yielded from the log-time responses, and vice versa. As such, important relationship between the responses might be lost. For instance, it is not possible to know whether an extremely easy stimulus (i.e., a stimulus that obtains a high proportion of correct responses) is as such because respondents tend to spend a high amount of time on it before giving a response or whether it also obtains fast responses. In the latter case, the stimulus can be considered as a good functioning one from both an accuracy and a time perspective. Similarly, if a stimulus has a low time intensity estimate (i.e., it obtains fast responses) combined with a low easiness estimate (i.e., it obtains a high proportion of incorrect responses), it should not be considered as a good functioning stimulus.

The separate modeling of accuracy and time responses assumes that the distributions of these variables are determined by different parameters, which are in turn generated by different processes (van der Linden, 2006). The accuracy and speed performance of one respondent

is constrained by a speed-accuracy trade-off. Once the speed-accuracy trade-off is set, the response time distribution of the respondent is solely determined by his/her speed. Similarly, the distribution of the accuracy responses only depends on the respondent's ability. However, when a population of respondents is considered, it is not possible to assume a single speed-accuracy trade-off, and a dependency between the accuracy and time responses should be expected (van der Linden, 2006, 2007). The relationship between the parameters governing the accuracy and speed performance can be understood at a second level of modeling, as illustrated in the hierarchical model by van der Linden (2007). As the name suggests, the hierarchical model posits two levels of modeling. At a first level, the accuracy and log-time responses are modeled separately. An IRT model is used for modeling accuracy responses, while the log-normal model is used for modeling the log-time responses. Each model yields stimuli and respondents' parameters explaining the accuracy and log-time responses. At a second level, two models are assumed to explain the relations between the respondents' parameters (i.e., *population model*) and the stimuli parameters (i.e., *item-domain model*). The population model assumes a multivariate normal distribution to describe the population from which the respondents are drawn. The multivariate distribution is defined by the respondents' parameters obtained from the accuracy and log-time models. The item-domain model describes the domain (population) of the items from which the items are drawn by assuming a multivariate normal distribution defined by the stimuli parameters obtained from the accuracy and log-time models.

Undoubtedly, the second level of modeling introduced by van der Linden (2007) would provide further insights on the functioning of implicit measures, concerning both the stimuli and the respondents. Nonetheless, Rome wasn't build in a day. As van der Linden (2006) himself did, the first step for a hierarchical approach is to find the appropriate models for the first level of modeling. Despite neither the Rasch model nor the log-normal model are breaking news in Psychometrics, their application to implicit measures data with the Linear Mixed-Effects model approach followed in this thesis is rather new. As such, we first wanted to find an appropriate and reliable approach to the separate modeling of implicit measures accuracy and time responses, able to be used as stand-alone models for each type of responses.

This thesis was mainly focused on the modeling of the IAT-family implicit measures, namely the IAT and the SC-IAT. Both the IAT and the SC-IAT are based on the accuracy and speed of the responses, and they exploit the logic of responses compatibility to sort different stimuli in contrasting conditions. The categorization happens by means of two response keys. Other implicit measures, such as the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), exploits the same logic of response compatibility but in favor of the inhibition of the responses in contrasting conditions. As such, only one response key is needed. In the GNAT, only two categories at the time are presented, such as *Coke* and *Good*. Along with the stimuli belonging to these target categories, stimuli representing either other beverages or negative attributes are presented. The task is to identify the stimuli belonging to the target categories by pressing the response key and to do nothing (i.e., inhibit the response) when the distractors appear on the screen. The same task has to be performed in a contrasting condition where *Coke* exemplars and *Bad* attributes are the reference categories. The underlying idea is that it would be easier to press the response key when the reference categories are strongly associated between each other than when they are not. The structure and the type of task characterizing the GNAT make it not possible to obtain a response time of the correct inhibition when a distractor is presented. Consequently, the scoring of the GNAT is entirely based on the accuracy responses. Given that the accuracy and log-time response models presented in this thesis do not rely on each other to be applied, the model based on accuracy responses can be used to model the accuracy responses of the GNAT to obtain a Rasch parametrization of the data. If the GNAT is administered with other implicit measures, the accuracy responses of both measures can be modeled together with a comprehensive modeling such as that presented in this thesis.

So far, the modeling framework introduced in this work has been applied with main purpose of validating it, for both sand-alone implicit measures and multiple measures administered together. However, a more practical application is missing. For instance, this approach might be used for assessing the features of the IAT administration procedure on the respondents' performance. While it is known that some of features of the IAT administration, such as the order of presentation of the associative blocks, do influence the respondents' perfor-

mance (e.g., Greenwald et al., 2003), the effect of other features, such as the presentation of a feedback, is less investigated. Following the LMMs approach of this thesis might be particularly useful for at least two reasons. First, it would allow to carry out the investigation in a latent variable modeling framework. Second, if the investigation of the effect of the administration features is carried out in a *within-subjects* experimental design, this approach allows for accounting for the dependencies of the observations. As such, it provides more reliable estimates, which in turn lead to more valid and generalizable inferences.

8.3 In the end

Despite the limitations, the results across the studies reported in this thesis highlighted the main aspects that have to be taken into account when analyzing IAT data.

Firstly, the consequences of not considering the fully-crossed structure of implicit measures and its related sources of variability and dependencies have been highlighted in terms of less reliable inferences of the constructs under investigation and a lower predictive ability of behavioral outcomes. The sensitivity of typical scoring methods to across-trials variability was blatant in the second empirical application of Chapter 5. The predictive ability of the D score was improved just by reducing the across-trials variability with the selection of some of the stimuli. One would have expected that the performance of the D score computed on the least informative stimuli would have led to a worse prediction than both the one computed on the entire data set and that computed on the reduced data set containing only highly informative stimuli. Conversely, the reduction of the across-trials variability was the feature that mostly impaired the reliability of the D score. Even the D score computed on the least informative stimuli showed a better predictive performance than the one computed on the entire data set (i.e., the one mostly affected by the across-trials variability).

Another feature of interest is related to the use of differential measures. Across studies, differential measures showed their inadequacy for expressing the implicit construct under investigation. Differential measures resulted in a lower predictive ability than that provided by the linear combination of their single components.

Regardless of the methodology used for analyzing implicit measures data, the predictive ability of implicit measures was always outperformed by that of explicit measures. This result might be due to the fact that the behavioral task was presented right after the questions on the explicit chocolate evaluation. Consequently, the preferred chocolate might have been made salient by the explicit questions, and the choice might have been made accordingly. Another explanation can be given by considering the nature of the assessment provided by implicit measures. Indeed, implicit measures are supposed to measure the tendency to associate target objects, like the two types of chocolate, with positive and negative attitudes. Clearly, a measure like that reflects the like/dislike towards the specific target object. It is not a measure of how much one (or both) the target objects are wanted. According to Meissner et al. (2019), this is the feature of implicit measures leading to their low predictive ability of behavioral outcomes. Indeed, the choice might be more driven by a *wanting component* (i.e., how much an object is desired) rather than a *liking component* (i.e., how much an object is positively or negatively evaluated), which is the measure obtained from implicit measures. Nonetheless, the explicit assessment on the chocolate preference asked specifically how much respondent liked dark and milk chocolate. Consequently, also explicit measures aimed at the liking component and not at the wanting one.

Appendix A

R code for estimating Rasch model and log-normal model from IAT and SC-IAT data.

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.

The example is based on the Coke-Pepsi IAT example of Chapter 1. The estimation of the Rasch model and log-normal model from the SC-IAT data follows the same procedure. Consequently, the illustration is solely based on IAT data, but it can easily be implemented on SC-IAT data without any further changes besides the name of the data set.

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 (SC-IAT) (factor with two levels such as `mappingA` and `mappingB`).
- `stimuli`: Column containing the labels identifying each stimulus (e.g., `good`, `bad`, `coke1`, `pepsi1`).
- `latency`: Column containing the latency of the IAT (SC-IAT) responses. Latency can

be expressed in seconds or milliseconds. In case the IAT (SC-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 (SC-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 reported in Chapter 5, 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 the SC-IATs reported in Chapter 7, respondents were presented with 72 trials in each associative condition, hence 144 observations (rows) for each respondent (in each SC-IAT) were obtained.

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 5.1 of Chapter 4) the estimates of the stimuli are centered at 0 (argument `(1|stimuli)`), while in Model 3 (Table 5.1 of Chapter 4) respondents estimates are centered at 0 (argument `(1|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, Mächler, 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.1 Accuracy models specification

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

A.1.1 Model estimation

Model 1: The between–subjects variability is specified as random intercepts (i.e., `(1|subject)`). The between–stimuli variability is specified as random intercepts (i.e., `(1|stimuli)`) as well.

```
a1 <- glmer(correct ~ 0 + condition + (1|stimuli) + (1|subject),
data = data, # IAT (SC-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 at 0 (i.e., `(1|subject)`). The within–stimuli between–conditions variability is specified as the random slopes of the stimuli in the conditions (i.e., `(0 + condition|stimuli)`).

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

Model 3: The between–stimuli variability is specified as random intercepts, centered at 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),
data = data,
```

```
family = "binomial")
summary(a3) # summary of the results
```

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.1.2 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
```

A.2 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 `a` to `t`.

A.2.1 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
```

Appendix B

R code for a comprehensive modeling of implicit measures.

This appendix presents the R code used for obtaining the Rasch model and the log-normal model estimates from (Generalized) Linear Mixed-Effect models from IAT and SC-IAT data according to the comprehensive modeling approach presented in Chapter 6.2

The example is based on the Coke-Pepsi IAT and the Coke SC-IAT presented in Chapter 1. For illustration purposes, a Pepsi SC-IAT is considered as well. The associative conditions of the Pepsi SC-IAT are the Pepsi-Good/Bad one (PG condition) and the Pepsi-Bad/Good one (PB condition).

The data set should contain the following variables:

- `subject`: Column containing the respondents' IDs (can be numeric, a factor, or a string, as long as it is unique for each respondent).
- `measure`: Column containing the labels that identify the three implicit measures (e.g., `iat`, `cokesciat`, `pepsiciat`). This variable should be a factor with three levels.
- `condition`: Column containing the labels of the six associative conditions of the three implicit measures (e.g., `CGPB` and `PGCB` for the IAT, `CG` and `CB` for the Coke SC-IAT, `PG` and `PB` for the Pepsi SC-IAT). This variable should be a factor with six

levels.

- `stimuli`: Column containing the labels identifying each stimulus (e.g., `good`, `bad`, `coke1`, `pepsi`).
- `latency`: Column containing the latency of the IAT responses. Latency can be expressed in seconds or milliseconds.
- `correct`: Column containing the accuracy of the 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 of each implicit measure 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 of each implicit measure. For instance, in the IAT reported in Chapter 7, respondents were presented with 60 trials in each associative condition. Each of the SC-IATs was composed by 72 trials in each associative condition. The total number of observations (rows) for each respondent was 408 (i.e., 120 IAT observations, 144 Dark SC-IAT observations and 144 Milk SC-IAT observations).

Regardless of the dependent variable (i.e., either accuracy or log-time responses), the first model is the Null model in which both respondents and stimuli are specified as random intercepts across conditions and across implicit measures. The fixed effect is the effect of the implicit measure. Since the fixed intercept is set at 0, the estimates for each level of the fixed effect can be considered as the marginal *log-odds* (accuracy models) or the marginal expected average log-time response (log-time) models.

In the second model, the multidimensionality of the implicit measure is allowed at the respondents' level while stimuli are centered at 0. In other words, the random slopes of the respondents in the implicit measures (`0 + measure|subject`) and the random intercepts of the stimuli (`1|stimuli`) are specified.

Finally, in the third model the multidimensionality of the associative condition of the specific implicit measure is allowed at the respondents' level, by specifying their random

slopes in the associative conditions of each measure (`0 + condition|respondent`). Stimuli are specified as random intercepts (`1|stimuli`).

B.1 Accuracy models specification

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

B.1.1 Model estimation

Model 1: The effect of the implicit measure is specified as a fixed effect. The between–subjects variability is specified as random intercepts (`(1|subject)`). The between–stimuli variability is specified as random intercepts (`(1|stimuli)`).

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

Model 2: The effect of the implicit measure is specified as a fixed effect. The between–stimuli variability is specified as random intercepts centered at 0 (`(1|stimuli)`). The within–subjects between–measures variability is specified as the random slopes of the subjects in the implicit measures (`(0 + measure|subject)`).

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

Model 3: The between–stimuli variability is specified as random intercepts, centered at 0 (`(1|stimuli)`). The within–subjects between-conditions variability is specified as the ran-

dom slopes of the respondents in the conditions of each implicit measure (`(0 + condition|subject)`). The fixed effect is the associative condition of each implicit measure.

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

B.1.2 Model comparison

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

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

Models 2 and 3 have the same degrees of freedom. As such, the χ^2 statistics resulting from their comparison is meaningless, and only comparative fit indexes should be used instead.

B.1.3 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)$subject), # Respondents' IDs
ability = coef(a1)$subject[, 1] # Select only the random
)                       # intercepts estimates
```

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 only the random
```

```
)                        # intercepts estimates
```

Model 2 results in measure–specific respondents' parameters and overall stimuli parameters. Respondents' measure–specific ability parameters can be extracted as follows:

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

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

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

Model 3 results in condition–specific respondents ability parameters and overall stimuli easiness parameters. Respondents' condition–specific ability parameters can be extracted as follows:

```
cond_ability <- coef(a3)$subject[, -1] # drop the first column
# (fixed intercepts set at 0)
```

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.2 Log-time models specification

The code for the estimation of the log-time models is the same as the one used for the estimation of the accuracy 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)`). Consistently, the code for extracting the log-normal model estimates

from the log-time models is the same as that used for extracting the Rasch model estimates from the accuracy models. For these reasons, only the code for the estimation of Model 3 and the related code for extracting the log-normal model estimates are 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 between log-time models, the same code as the one used for accuracy models comparison can be employed. The names of the objects containing the models have to be changed accordingly, in this case from `a` to `t`.

B.2.1 Log-normal model parameters

The code for extracting the log-normal model parameters from log-time Model 3 is reported. The same code used for extracting the parameters for the accuracy models can be employed 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 interceptse set at 0)
```

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
```

References

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory*. Singapore, Singapore: Springer.

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

Attali, D. (2018). *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. Retrieved from `https://CRAN.R-project.org/package=shinyjs` (R package version 1.0)

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Bambini, V., & Trevisan, M. (2012). Un'interfaccia web per ricerche sul corpus e lessico di frequenza dell'italiano scritto. *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore*.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. doi: 10.3758/s13428-013-0410-6

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The Sorting Paired Features Task: A measure of association strengths. *Experimental Psychology*, *56*(5), 329–343. doi: 10.1027/1618-3169.56.5.329

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *44*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i–8. doi: 10.1002/j.2333-8504.1981.tb01255.x

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & R. Novick (Eds.), *Statistical theories of mental test scores.* Reading:MA: Addison-Wesley Publishing.

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

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(5), e1501. doi: 10.1002/wcs.1501

Bulmer, M., & Izuma, K. (2018). Implicit and explicit attitudes toward sex and romance in asexuals. *The Journal of Sex Research*, *55*(8), 962–974. doi: 10.1080/00224499.2017.1303438

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The "Big Five Questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences*, *15*(3), 281–288. doi: 10.1016/0191-8869(93)90218-R

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *shiny: Web Application Framework for R*. Retrieved from `https://CRAN.R-project.org/package=`

`shiny` (R package version 1.2.0)

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

Costantini, G. (2018). *Implicit Association Test Scores Using Robust Statistics*. Retrieved from `https://CRAN.R-project.org/package=IATscores` (R package version 0.2.1)

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.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5. doi: 10.1037/0022-3514.56.1.5

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

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62.

Ellithorpe, M. E., Ewoldsen, D. R., & Velez, J. A. (2015). Preparation and analyses of implicit attitude measures: Challenges, pitfalls, and recommendations. *Communication Methods and Measures*, *9*(4), 233–252. doi: 10.1080/19312458.2015.1096330

Epifania, O. M., Anselmi, P., & Robusto, E. (2019). Dscoreapp: An user-friendly web application for computing the implicit association test d-score. *Journal of Open Source Software*, *4*(42), 1764. doi: 10.21105/joss.01764

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

Epifania, O. M., Robusto, E., & Anselmi, P. (2020a). Filling the attitude-behavior gap: A Rasch modeling of the Implicit Association Test. *Journal of Experimental Psychology: General*. (submitted)

Epifania, O. M., Robusto, E., & Anselmi, P. (2020b, 2). Implicit social cognition through years: The Implicit Association Test at age 21. Retrieved from `https://advance.sagepub.com/articles/preprint/Implicit_social_cognition_through_years_The_Implicit_Association_Test_at_age_21/11914416` doi: 10.31124/advance.11914416.v1

Epifania, O. M., Robusto, E., & Anselmi, P. (2020c). Rasch gone mixed: A mixed model approach to the Implicit Association Test. *Testing, Psychometrics, Methodology in Applied Psychology*. (in press)

Faraway, J. J. (2016). *Extending the linear model with R.* (2nd ed.). Boca Raton; Florida: CRC Press.

Fatfouta, R., & Schröder-Abé, M. (2018). Agentic to the core? Facets of narcissism and positive implicit self-views in the agentic domain. *Journal of Research in Personality*, *74*, 78–82. doi: 10.1016/j.jrp.2018.02.006

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. doi: 10.1146/annurev.psych.54.101601.145225

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Asso-

ciation Test (IAT). *European Review of Social Psychology*, *17*(1), 74–147. doi: 10.1080/10463280600681248

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. doi: 10.1037/0022-3514.82.6.878

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. doi: 10.1177/0146167216684131

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

Glashouwer, K. A., Vroling, M. S., de Jong, P. J., Lange, W.-G., & de Keijser, J. (2013). Low implicit self-esteem and dysfunctional automatic associations in social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 262–270. doi: 10.1016/j.jbtep.2012.11.005

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27.

Greenwald, A. G., & Banaji, M. R. (2017). The Implicit Revolution: Reconceiving the Relation Between Conscious and Unconscious. *American Psychologist*, *72*(9), 861-871. doi: 10.1037/amp0000238

Greenwald, A. G., & Lai, C. K. (2020). Implicit Social Cognition. *Annual Review of Psychology*, *71*. doi: 10.1146/annurev-psych-010419-050837

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 Soclal 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), 419-445. doi: 10.1037/a0015575

Hiller, T. S., Steffens, M. C., Ritter, V., & Stangier, U. (2017). On the context dependency of implicit self-esteem in social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *57*, 118–125. doi: 10.1016/j.jbtep.2017.05.005

Houwer, J. D., & Hermans, D. (1994). Differences in the affective processing of words and pictures. *Cognition & Emotion*, *8*(1), 1–20. doi: 10.1080/02699939408408925

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

Klein, M., Weksler, N., Gidron, Y., Heldman, E., Gurski, E., Smith, O. R. F., & Gurman, G. M. (2012). Do waking salivary cortisol levels correlate with anesthesiologist's job involvement? *Journal of Clinical Monitoring and Computing*, *26*(6), 407–413. doi: 10.1007/s10877-012-9367-8

Linacre, J. M. (1989). *Many-Facet Rasch measurement*. Chicago, IL: MESA Press.

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

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. New York, NY: Routledge.

Martin, D. (2016). *IAT: Cleaning and Visualizing Implicit Association Test (IAT) Data*. Retrieved from `https://CRAN.R-project.org/package=IAT` (R package version 0.3)

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

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

Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69.

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

Pastore, M. (2015). *Analisi dei dati in psicologia*. Il mulino.

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

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. New York; New York: Springer Science & Business Media.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna,

Austria. Retrieved from https://www.R-project.org/

Raaijmakers, J. G. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*(3), 141. doi: 10.1037/h0087421

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416–426. doi: 10.1006/jmla.1999.2650

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

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

Software, C. (2011). *Inquisit 3.0.6.0*. https://www.millisecond.com.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental psychology*, *56*(4), 283–294. doi: 10.1027/1618-3169.56.4.283

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

Storage, D. (2018a). *IATanalytics: Compute Effect Sizes and Reliability for Implicit Association Test (IAT) Data*. Retrieved from https://CRAN.R-project.org/package=IATanalytics (R package version 0.1.1)

Storage, D. (2018b). *IATScore: Scoring Algorithm for the Implicit Association Test (IAT)*. Retrieved from https://CRAN.R-project.org/package=IATScore (R package version 0.1.1)

Thissen, D. (1983). Timed testing: An approach using item response theory. In *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). New York; New York: Elsevier.

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

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

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from `https://ggplot2.tidyverse.org`

Wigboldus, D. H. J., Holland, R. W., & van Knippenberg, A. (2004). Single target implicit associations. *Unpublished manuscript*.

Wilkie, J. E., & Bodenhausen, G. V. (2015). The numerology of gender: Gendered perceptions of even and odd numbers. *Frontiers in psychology*, *6*, 810. doi: 10.3389/fpsyg.2015.00810

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101. doi: 10.1037//0033-295X.107.1.101

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

Wright, B. D. (1997). A history of social science measurement. *Educational measurement: Issues and practice*, *16*(4), 33–45.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago; Illinois: MESA Press.

Zayas, V., & Shoda, Y. (2005). Do automatic reactions elicited by thoughts of romantic partner, mother, and self relate to adult romantic attachment? *Personality and Social Psychology Bulletin*, *31*(8), 1011–1025. doi: 10.1177/0146167204274100

Zogmaister, C., & Castelli, L. (2006). La misurazione di costrutti impliciti attraverso l'Implicit Association Test. *Psicologia Sociale*, *1*, 65–94.