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 Models approach for modeling implicit measures within a Rasch framework

Advisor: Prof. Egidio Robusto
Co-Advicor: 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

Contents

Preface **v**

1 Theories, Processes, Measurement **1**

1.1 Automatic and controlled processes 2

1.2 The Implicit Association Test 4

1.3 Fields of application 5

1.4 The Single-Category Implicit Association Test 10

1.5 Fully-crossed design 13

1.5.1 More than one implicit measure 15

2 Classic scoring of implicit measures **17**

2.1 IAT D-score 18

2.2 SC-IAT D-score 20

2.3 A fairer comparison 20

2.3.1 Method 22

2.3.2 Results 26

2.3.3 Discussion and Conclusions 34

2.4 R development 37

2.4.1 `implicitMeasures` 37

2.4.2 DscoreApp 40

3 Formal modeling **45**

3.1 Multionomial Models 46
3.1.1 Quad Model. 46
3.1.2 ReAL Model 47
3.2 Time and Accuracy models 48
3.2.1 Diffusion Model 48
3.2.2 Discrimination-Association Model 50
3.3 Rasch Modeling 51

4 Rasch model model and log-normal model 53
4.1 Rasch Model 54
4.2 Log-normal Model 54
4.3 Linear Mixed Effects Models 55
4.3.1 (Generalized) Linear Model and Rasch Model 56
4.3.2 Extending to Multidimensional data 57

5 Models specification 59
5.1 Maximal Model 60
5.2 Models random structure 62

6 Empirical application to IAT data 67
6.1 Rasch gone mixed 68
6.1.1 Method 68
6.1.2 Race IAT: Results 69
6.2 Filling the Gap 73
6.2.1 Method 73
6.2.2 Chocolate IAT: Results 74
6.3 Discussion and Conclusions 79

7 Make it comprehensive 83

8 The feedback effect on implicit measures performance 85

9 General discussion **87**

Last note **89**

References **91**

Preface

The advent of measures able to infer mental processes from the speed of respondents to computerized categorization tasks opened the access to the assessment of processes that lie beyond people's awareness. Despite they are outside of awareness, these processes influence people's attitudes, preferences, and behaviors towards different objects. They can be captured by measures that are specifically deigned for tapping into them, measures that are know as "implicit measures". The use of implicit measures became more and more popular in social sciences, also thanks to the availability of more and more software for the administration of computerized categorization tasks. Implicit measures received a lot of positive and negative attention throughout the past two decades, and they became extremely popular in social sciences. However, a lot of work needs to be done to find a psychometrically sound approach to their modeling.

Usually, implicit measures are scored by averaging across stimuli to obtain respondents-specific scores to be used 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 for data set obtained with typical Social psychology methodologies, such as implicit measures, is of the uttermost importance.

The main objective of the Thesis is to provide new methods for more rigorous analyses of implicit measures data. In the long term, the repercussion of more rigorous data analyses can be observed in the replicability of the obtained 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 measures scores (easy path).

The sound path, which is also the main one, is an attempt at finding new approaches for the analysis of implicit measures. This is done by combining a classic of Psychometric Theories, the Rasch model, with a Linear Mixed Effects Models 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 (IAT; Karpinski & Steinman, 2006). Accuracy and time responses of implicit measures are modeled separately with distinct models. Consequently, the parameters are either explaining the processes leading to the accuracy responses or those leading to the time responses. The relationship linking these parameters can be explained and understood in a second level modeling (Van Der Linden, 2009).

Traditionally, Item Response Theory and Rasch modeling treats 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. Indeed, the fully-crossed design characterizing the IAT (see Chapter 1.5) allow one to conceptualize the stimuli as a manifestation of the super-ordered category they represent. Consequently, they are just one the possible set of stimuli that can be drawn from the same population of stimuli. Following this line of reasoning, it made more sense to consider them as random factors, and to treat them as random effects. Besides being a statistically more sound approach, acknowledging the sampling variability of the stimuli implies that each stimulus potentially has a 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 for gathering all the information they convey (Wolsiefer, Westfall, & Judd, 2017). Considering both respondents and stimuli as sample drawn from larger population and hence treating them as random factors allows for obtaining detailed and generalizable information not only at the respondents level, but also to consider, investigate, the functioning of each stimulus and its impact to the observed responses.

This approach can be followed also when multiple implicit measures are administered concurrently, which is a quite common practice in experimental psychology. In such cases, new sources of variability are added to the one already present in each implicit measure. 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, using Linear Mixed Effects Models for the conjoint analysis of multiple implicit measures within a Rasch framework is not a common approach to implicit measures data. 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 predictive capacity of behavioral outcomes. The approach followed in the fair path is more in line with the typical analyses performed on implicit measures data. Size effect measures are the most popular and used scoring procedures for the IAT and the SC-IAT, and they are often used for comparing their performance 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, one of which is the lack of control on the sources of random variability in the data. Furthermore, both the scoring and the 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, ending up in misleading results. New scoring algorithms for the IAT and the SC-IAT are introduced in the attempt of minimizing not necessary procedural differences potentially affecting the comparison. The computation procedure of effect size measures itself cannot overcome the issues of the sources of random variability characterizing implicit measures data. By aligning the differences in the scoring procedure of the IAT and the SC-IAT, the new scoring alternatives should at least provide a means for a fairer comparison between the IAT and the SC-IAT. Consequently, the new, aligned, scoring algorithms should produce 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 will always end in the same results, and the results can be easily and openly replicated. In the long term, this would help for the replicability of the results obtained through implicit measures.

In the first Chapter, 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.

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. Other implicit measures have been introduced with the aim of providing an absolute measure of a target object. The SC-IAT (Karpinski & Steinman, 2006) is a direct modification of the IAT procedure, where one of the target objects is dropped. It is often used as an alternative to the IAT when the aim is to obtain an absolute measure towards one object. The description of the SC-IAT is provided in Chapter **??**. 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 it it is not correctly accounted for.

Fair and easy paths are both presented in Chapter **??**. The typical and new, modified scoring procedures of the IAT and the SC-IAT are illustrated. The alignment of both administration and scoring procedures of the IAT and the SC-IAT in the new scoring algorithms, should provide a comparison between the predictive ability of the two measures more centered on the implicit measures themselves than on the scoring/administration avoidable differences. The comparison is usually based on the implicit measures predictive ability of behavioral outcomes, and the IAT tends to outperform the SC-IAT. The rationale for the introduction of the new scoring algorithms is that, if by aligning the administration and scoring procedures of the two implicit measures, while acknowledging their main features, the IAT still outperforms the SC-IAT, its better performance should be ascribable to the measure itself, and not to procedural artifacts. The results of an empirical study where typical and modified scoring procedures were compared are reported. Regardless of the scoring algorithms, the measure obtained from the IAT always outperformed the one obtained from the SC-IAT. Limitations of these results might be related to the choice task considered as a behavioral outcome, as further illustrated in the Chapter.

Regarding the easy path component of Chapter **??**, two open source alternatives for the computation of the IAT and the SC-IAT typical scoring procedures are illustrated. One of them is a Shiny app (DscoreApp) 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 (`implicitMeasures`). 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. Beyond making the computation easier, it also guarantees for the replicability of the results. Indeed, researchers often fail to report the specific *D* score algorithm they have used for scoring the IAT. Additionally, replicability of the results is undermined by the high number of 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.

Among others, DscoreApp presents two main shortcomings. One of them is an intrinsic limitation of Shiny apps. By putting the code into the shiny interface, it is not possible to call it and run it from the command line, hence making it impossible to reproduce. However, 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.

`implicitMeasures` 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 the individual level or sample level.

Chapter **??** 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. Quad model (Conrey, Gawronski, Sherman, Hugenberg, & Groom, 2005) and ReAL model (Meissner & Rothermund, 2013) are based on accuracy responses, while Diffusion model (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007) and 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 from the most controlled processes intervening during the performance at the IAT. A common finding is that automatic associations are just one of the possible processes intervening during the performance at the IAT, and that other controlled processes, such as recoding the stimuli (ReAL, Diffusion Model) or suppressing 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, the fully-crossed structure of the IAT is overlooked.

Rasch modeling of 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. The applications of the Rasch model to IAT data that have been done 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 Rsach modeling does not account for the random noise in the data. The random noise due to the different sources of variability in the IAT data 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 **??**. Its similarities with Generalized Linear (Mixed Effects) Models and its limitations when it comes to its application to complex data structures, such as that of the IAT, are presented as well. Given that 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 dependency in 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 trials levels, hence resulting in more reliable estimates of the model parameters.

The log-normal model is introduced in the first section of Chapter **??** 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 **??**) 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, hence the sources of dependency are addressed.

Rasch model and log-normal estimates are not directly resulting 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 different degrees of details on either the respondents or the stimuli.

The second section of Chapter **??** presents the specification of models with different random structures for a meaningful Rasch and log-normal analysis of IAT data. Three models for accuracy responses and three models for log-time responses are specified for obtaining Rasch model and log-normal estimates, respectively. Besides the assumption on the distribution of the error term, the random structures of the accuracy and 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 differentiate themselves 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 **??** are illustrated in Chapter **??**. The first application was aimed at investigation of the validity of the proposed model for the analysis of IAT data. To pursue this aim, a Race IAT was employed and the relationship between the estimates obtained from Rasch and 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 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, and, consequently, lead to a better prediction of behavioral outcomes, than the one given by the typical scoring procedure. The second application was also aimed at testing the usefulness of the condition–specific stimuli estimates. If the stimuli estimates truly allows for pinpointing the most informative, as well as the least informative stimuli, a higher amount of information should be obtained by just selecting the most informative stimuli. Therefore, a data set with a fewer number of stimuli, and hence, of trials is obtained. In such data set, the across trials variability due to heterogeneity of the stimuli should be reduced. Consequently, 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.

Rasch model and log-normal estimates did result in a better inference of the implicit preference that allowed for 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 reducing the across-trials variability, and 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. Grounding on these results, it is possible to speculate that, no matter of the information provided by the stimuli, the across-trials is what mostly biases the *D* score computation, and just by reducing it, it is possible to obtain better measures more related to external variables.

In Chapter **??**, 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. 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 **??** has a main, outstanding fallacy, that is, the *post hoc* separation of implicit measures administered concurrently to the same respondents.

The issues related to the data structure of implicit measures administered alone have already been highlighted in Chapter **??**. When multiple measures are administered concurrently, each of them comes with its peculiar data structure and its method variance. 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 might be present. 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 **??** 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 **??** 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 **??** is presented in Chapter **??**. Data are the same as those in Chapter **??**, hence including one IAT and two SC-IATs. Implicit measures have been modeled separately with the (G)LMMS of Chapter **??** for obtaining Rasch model and log-normal estimates from each of them individually. This was done for mainly two reasons. Firstly, to investigate the soundness of the proposed approach for modeling other measures 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 is 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 more reliable than the *D* score, resulting in a better prediction of the choice. 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 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.

Finally, Chapter 9 summarizes the findings of all other chapters, presents the limitation of the approach, and introduces future directions.

Chapter 1

Implicit measures: Between measurement, theories, and cognitive processes

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) is then introduced as well as an overview of its use from the year of its introduction to current days, and its main fields of application. Then, 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). The chapter ends with the illustration of the data structure resulting from the IAT and the SC-IAT administration.

1.1 Automatic and controlled processes

Throughout the past decades, the interest in the implicit investigation of attitudes and preferences has been constantly growing among social scientists. 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).

The peculiarity of implicit processes is that they can be activated by a triggering stimulus, but they can remain outside of people's awareness. Despite the lack of awareness, implicit processes can still guide behaviors and decisions (e.g., Greenwald & Banaji, 1995), as well as social judgments and behaviors (Greenwald & Lai, 2020).

According to dual-process theory of prejudice and social judgments (Devine, 1989), both implicit (or automatic) processes are concurrently intervening in shaping how people behave in their social environment. Differently for implicit processes, which are automatically activated by triggering stimuli and remain outside of people's consciousness, controlled processes are characterized by their controllability. Besides, controlled processes need for the availability of cognitive resources to be activated. Finally, since controlled processes can be controlled by individuals, they are more susceptible to faking in socially sensitive contexts than implicit processes.

A common trend for the assessment of the controlled and automatic components of psychological constructs is to administer a measure tapping the former one, like a self-report scale, concurrently with a measure tapping the implicit component. Controlled processes, as assessed by self-report, have been found to be highly correlated with implicit measures investigating the same construct (e.g., Nosek, 2007). However, this result does not hold when the assessment involves socially sensitive topics, such as racial prejudice (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Nosek, 2007). The distinction between the different influences of controlled and implicit processes can also be seen in their abilities to predict different type of behaviors. Controlled processes have been found to have a higher probability of predicting a deliberate behavior, while implicit processes have been found to be particularly effective in predicting non-deliberate and rather automatic behaviors (e.g., Wilson, Lindsey, & Schooler, 2000; Perugini, 2005).

Different theoretical models have been introduced to explain the apparent dissociation between automatic (or implicit) and controlled (explicit) processes (Perugini, 2005).

According to some authors (e.g., Fazio & Olson, 2003), implicit and explicit measures of psychological constructs should be considered as deriving from a single–representation of the specif construct under investigation, and not as deriving from distinct and unrelated representations. Consequently, both processes are giving a unique contribution in the prediction of the behaviors, and in some cases, one of the attitudes over the other one has the predictive power.

In the Model of Dual Attitudes (Wilson et al., 2000), the coexistence in memory of two attitudes (i.e., different evaluations of the same object) is hypothesized. Corresponding to the combination of awareness of the implicit attitude and the availability of cognitive resources needed by the explicit attitude to be activated, one will override the other. Implicit attitudes are going to solely influence implicit responses and/or non-deliberative behaviors, or, in other words, responses or behaviors that cannot be controlled. Conversely, explicit attitudes are going to influence only explicit responses and/or deliberative behaviors.These responses/behaviors are under conscious control, and ca be understood as the expression of the relevant explicit attitude.

Several paradigms have been introduced fro tapping the implicit component od attitudes, such as the Implicit Association Test (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 (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 (Epifania, Robusto, & Anselmi, 2020; Greenwald et al., 2009; Greenwald & Lai, 2020).

1.2 The Implicit Association Test

The IAT measures the strength of 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 exemplars 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 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.

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

Block	Trials	Function	Left Response key	Right Response key
B1	20	Pure practice	Good words	Bad words
B2	20	Pure practice	Coke	Pepsi
B3	20	Associative practice	Good & Coke	Bad & Pepsi
B4	40	Associative test	Good & Coke	Bad & Pepsi
B5	20	Pure practice	Pepsi	Coke
B6	20	Associative practice	Good & Pepsi	Bad & Coke
B7	40	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 usual structure of an IAT (illustrated in Table 1.1) is composed of 7 blocks. The first two blocks are pure practice blocks in which the respondents have to sort the stimuli belonging to either the evaluative dimension categories (Block B1) or the stimuli belonging to the object categories (Block B2). These blocks have the purpose of letting the respondents familiarize with both 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). In Block B5, the locations for the labels of the object categories switch their position, and this block help respondents to familiarize with the new labels locations. Block B6 and B7 constitutes 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).

During the IAT administration, respondents might be given feedback of their performance. Specifically, 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 responses. 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).

1.3 Fields of application

A recent literature review (Epifania et al., 2020) showed an increasing use of the IAT in wider and more varied fields of applications. Part of the results of the literature review are reported here. 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 1.1 depicts the trend lines for each filed of application from 1998 to 2019.

The dashed line in Figure (1.1) represents the average trend of the IAT use across different fields of applications, and it clearly point a constant an on-going growth in the IAT use

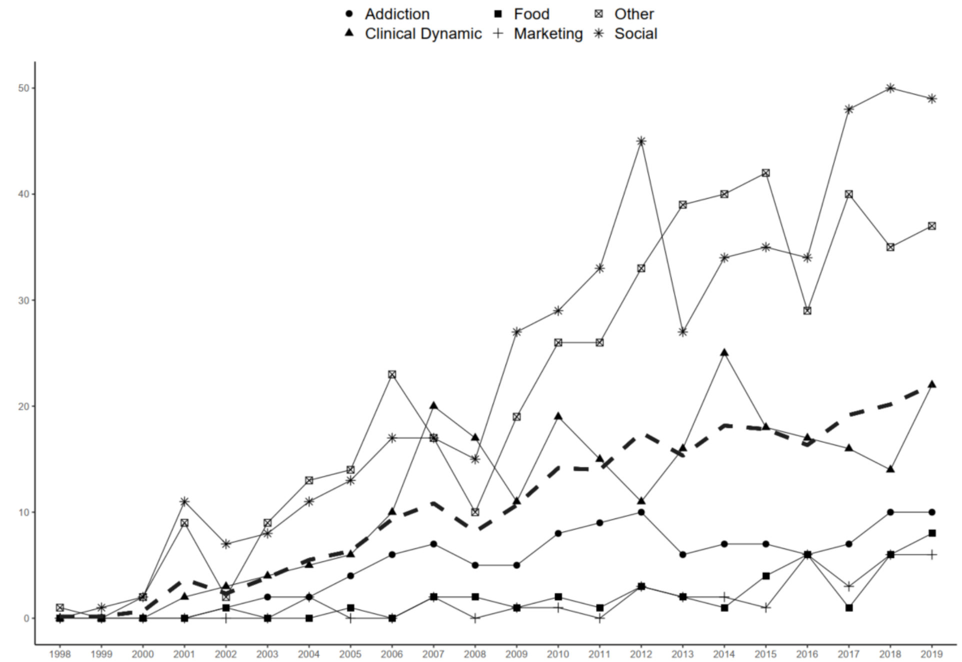


Figure 1.1: Trend lines of each field of application of the IAT throughout from 1998 to 2019.

throughout the years.

Trend lines for the fields of application Social psychology and Other appear always above the mean trend line. Clinical and dynamic psychology trend line is the most inconsistent one throughout the years. Specifically, it shows a steady growth until 2007-2008, with a drop in 2009, followed by a vacillating trend until 2015, where it encounters a sort of *plateau* with a final peak in 2019. The trend lines of fields pf applications 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.

Figure 1.2 Reports the most common topics for which the IAT has been used throughout the years.

It is not surprising to note that the most common use of the IAT was for the investigation of implicit stereotypes and attitudes, such as implicit racial prejudice, gender stereotypes, attitudes towards obese people and other social groups, such as psychiatric patients and people with disabilities. Indeed, the IAT was specifically designed for tapping the automatic aspects

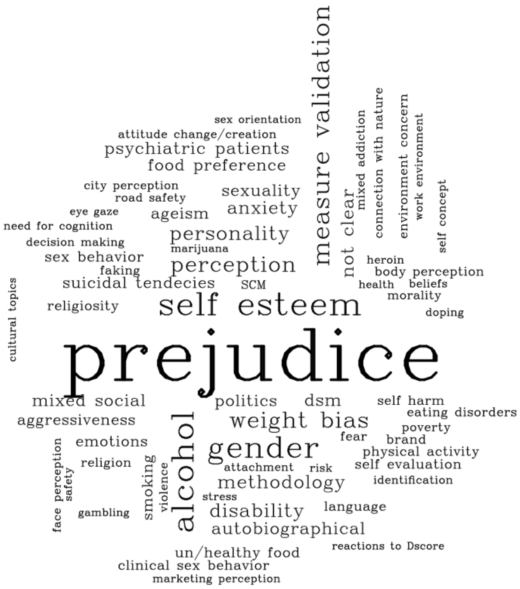


Figure 1.2: IAT fields of application word cloud. The bigger the word font, the more common the use of the IAT for the investigation of that specific topic.

of attitudes, making it more resistant to self-presentation biases (e.g., Greenwald et al., 2009). As such, its use is particularly appealing for the investigation of socially sensitive topics.

Other interesting application of the IAT concerns the investigation of addiction towards different drugs or unhealthy behaviors, such as alcohol consumption or smoking, personality traits, and self-perception.

In the following paragraphs, a brief summary of the IAT applications in each of the above-mentioned fields is provided.

IAT in Social psychology. Given the IAT resistance to self-presentation strategies and hence to responses altered by social desirability (e.g., Greenwald et al., 2009), its use is particularly popular for the investigation of socially sensitive topics. Most of the studies employing the IAT in Social psychology deals with the investigation of out-group stereotypes and prejudice, specifically related to racial prejudice. Studies on this topic are followed by studies focused on gender stereotypes, and on the investigation of illness-related attitudes. Illness-related attitudes indicate 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, 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 multi 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.

It is interesting to note the specific samples on which studies using the IAT have been carried out. Specifically, the IAT appears to be widely used among samples composed of hospital personnel (i.e., doctors, nurses, therapists), adults, undergraduates, and work personnel (i.e., CEOs, hiring personnel, university boards).

The common goal of the studies investigating out-group prejudice and stereotypes, illness related attitudes, and weight bias among hospital personnel was to understand whether the implicit bias towards the specific out-group influenced the quality of the provided care on the probability of suggesting a specific surgery. The IAT was applied in a similar vein on samples composed of work personnel. Specifically, these studies were aimed at understanding whether implicit biases could undermine the probability of people belonging to minority groups, such as African-American people in the US, to be hired in a corporation or to win a position in Academia. Another interesting IAT application was for the investigation of the implicit out-group prejudice or implicit gender stereotypes of teachers. The aim of these studies was to understand whether teachers' gender implicit attitudes (e.g., implicitly associating males with math and females with humanistic subjects) or their racial implicit attitudes (e.g., associating European-American children with higher intelligence than Hispanic children) could influence their behaviors towards their pupils.

IAT in Clinical and dynamic psychology. Concerning 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 DSM psychological disorders (both personality and mood disorders) 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. Vast majority of studies using the IAT in research on addiction 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 and smoking or drinking and gambling, was quite uncommon. The IAT was used for investigating cocaine addiction in just one study. These studies are mostly carried out on clinical samples of people with a diagnosis of addiction to some substance, but also studies carried out on college students are fairly common.

IAT in Food research. The most common aim with which the IAT is used in Food research was to investigate the preference for different kind of food, the preference for healthy over unhealthy food, or food perception. IAT was also employed for investigating attitudes towards dieting. Less common topic 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 field of applications of the IAT. Most of the studies are focused on the implicit evaluation of the preference for a brand over another. 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. Studies included in this field of application include 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 the 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 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, 1960) 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.

1.4 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 the contrasted one. Many attitudes or preference objects present a clear contrasted category, so that it is meaningful and useful to investigate the relative preference/attitudes (Greenwald & Farnham, 2000). 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 (faster responses in sorting *Pepsi* images and *Bad* attributes with the same response key), or even a combination of the two. The resulting IAT effect is not able to actually acknowledge for the specific preference/dislike driving the responses. One could try to decompose the IAT effect by computing separate scores for the trials in which *Good* attributes and *Coke* images are associated and in which *Bad* attributes and *Coke* 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.

Clearly, when the focus of the research is on the assessment of the absolute positive or negative evaluation of a single object, the IAT might not represent the best choice. Consider the case of self-esteem evaluation. 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).

Moreover, since the final effect would depend 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. Consider the Coke-Pepsi IAT example again. The positive or negative evaluation of Coke is strongly dependent on the positive or negative evaluation of Pepsi. By changing the contrasted category, for example from Pepsi to Dr. Pepper, the evaluation of Coke might also change since this evaluation is relative to the alleged opposite of Coke.

Different alternative measures have been introduced to overcome the issue of the relativeness of the implicit measure, like the Single Category Implicit Association Test (SC-IAT; Karpinski & Steinman, 2006), which results from a slight modification of the IAT procedure. As the IAT, the SC-IAT measures the strength of associations between concepts by considering the speed and accuracy with which different stimuli are sorted in their reference category. The assumption that underlies the functioning of the SC-IAT is the same as that underlying the IAT functioning, 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 attributes categories, are presented. The usual structure of a Coke SC-IAT is illustrated in Table 1.2.

The SC-IAT is usually composed of 4 blocks, two of which (Blocks B1 and B3) are associative practice blocks used to let the respondents familiarize with both the stimuli and the

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

Block	Trials	Function	Left Response key	Right Response key
B1	24	Associative practice	Good & Coke	Bad
B2	72	Associative Test	Good & Coke	Bad
B3	24	Associative practice	Good	Bad & Coke
B4	72	Associative Test	Good	Bad & Coke

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

task. In the first critical block, that is, the first critical condition, the target object *Coke* and *Good* attributes share the same response key, while *Bad* attributes are sorted with the opposite response key (Coke/Good-Bad condition). In the contrasting associative condition, 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). The SC-IAT administration procedure usually includes a response time window (rtw) at 1,500 ms, after which the stimulus disappears and a warning message (e.g., "Respond more quickly!") is given to the respondent. Moreover, 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 presence of the rtw and the 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 *D-score* (Karpinski & Steinman, 2006, see Chapter 2.2).

1.5 Implicit measures fully-crossed design

In tasks like the IAT and the SC-IAT, stimuli representing each of the reference categories are presented multiple times, both within and between the associative conditions. Each stimulus exemplar is also presented multiple times to each respondent, both within and between the associative conditions. Thus stimuli and respondents are crossed with each other (multiple responses to the same stimulus by the same participant), and they are in turn crossed with the associative conditions (multiple responses on the same stimulus by the same respondents both within and between the conditions). Besides being the design characterizing both the IAT and the SC-IAT, this type of design (i.e, fully-crossed design) is rather common in experimental psychology, from linguistic to social cognition (Westfall, Kenny, & Judd, 2014). Data deriving from this design should be carefully analyzed to account for the sources of dependencies that might end into misleading results and undermine their replicability.

The multiple observations on the same respondents and on the same stimulus generate a random noise at the level of the single observation that cannot be ignored to obtain reliable parameters estimates. The fully-crossed structure results in a unique combination respondent × stimulus for every repetition of the stimuli in each associative condition. Consequently, different sources of variability and dependency – related to participants, stimuli, conditions, and to their interaction – have to be expected (Barr, Levy, Scheepers, & Tily, 2013; Westfall et al., 2014; Wolsiefer et al., 2017).

Suppose that two respondents, Ben and Angela, are administered with the Coke-Pepsi IAT (or the Coke SC-IAT) presented in Section 1.2 (or Section 1.4). Ben might be, on average, more accurate (or faster) than Angela, and this may be due to different reasons pertaining Ben's and Angela's characteristics. This between–response variability can be considered as the expression of the variability due to respondents' individual differences, across the associative conditions.

The exemplars used for representing the categories are important as well. Consider the *Coke* category. The logo of the company can be immediately recognized as belonging to the *Coke* category, hence having a higher probability of being immediately (and accurately sorted) into the correct category. Conversely, an old-fashioned image of a can of Coke might be less familiar and could hence require more time for being recognized and accurately sorted in the correct category. The difference caused by the properties of the stimuli generates a variability at the stimuli level (i.e., between stimuli variability), which can be considered as the expression of the variability due to the stimuli characteristics eliciting different responses, across the associative conditions.

Ben and Angela's performance might be impaired or facilitated by the associative conditions. For instance, the difference between their performance might be reduced in the condition in which Coke is associated with positive attributes and Pepsi is associated with negative attributes. In this case, Angela might be facilitated because she is particularly fond of Coke and it is easier for her to associate Coke exemplars with positive attributes, while Ben is already a good respondent. In the opposite condition, the difference in their accuracy or speed performance might be exacerbated. Angela, who is already a low accuracy respondents, might find particularly difficult to associate Coke with negative words. This variability (i.e., the within–respondents between conditions variability) can be considered as the effect of the IAT associative conditions on the respondents' performance. It expresses the deviation of each participant from the estimate of the average performance in each condition.

Also the categorization of the stimuli can benefit from the associative condition in which they are presented. For instance, the Coke exemplars might receive more accurate and faster responses when they share the response key with positive attributes than when they share it with the negative response key. This within–stimuli between–conditions variability can account for this intra-stimulus variability, and can hence be interpreted as the expression of how the IAT associative condition is affecting the stimuli categorization.

When these sources of variability are not included into a statistical model able to account for them, they can influence the estimation of any parameters, or the computation of any scores. Thus, ignoring the sources of variability and the dependency between the IAT or the SC-IAT observations may result in biased parameter estimates (such as incorrect standard errors), which may lead to an overestimation (or underestimation) of the importance of experimental effects (Judd, Westfall, & Kenny, 2017; McCullagh & Nelder, 1989). Moreover, by overlooking the sampling variation related to the stimuli, the information that can be gathered from them is completely neglected (Wolsiefer et al., 2017).

1.5.1 More than one implicit measure

It is not uncommon that the IAT and the SC-IAT are used concurrently to obtain both a comparative measure between the two objects of interest and an absolute evaluation towards each of them (e.g., Bulmer & Izuma, 2018; Glashouwer, Vroling, de Jong, Lange, & de Keijser, 2013). When both the implicit measures are used, their data are analyzed separately, and their effects are usually expressed by separate *D-score*s. The *D-score*s are effect size measures obtained by standardizing the difference in the average response times between the two contrasting conditions. These indexes are easy to compute and can provide a quick and interpretable measure of the implicit constructs under investigation. Nonetheless, they come with issues that cannot be ignored, both when the IAT and the SC-IAT are used as single measures and when they are concurrently employed.

Beyond the issues related to the computation of scores like the *D-score* on data structure such as those of the IAT and the SC-IAT, there are also issues arising from the concurrent use of the implicit measures. When the IAT and the SC-IAT are used together and separate *D-score*s are computed, the within–respondents between–measures variability is completely ignored, as well as method specific variance. Any possible carry-over effects or dependency due to the multiple observations on the same respondent, both within and between measures and their respective associative conditions, are not accounted for. Furthermore, since the same stimuli are usually employed across the IAT and the SC-IAT, other sources of dependency at the stimuli level have to be expected. Consequently, the resulting *D-score*s are including a proportion of unwanted error variance directly related to the sources of dependency at different levels.

Chapter 2

Classic scoring of implicit measures

This chapter deals with the classic scoring measures for the IAT and the SC-IAT. Firstly, the scoring algorithms for both the IAT and the SC-IAT are illustrated and described. Then, new algorithms for the scoring of both implicit measures are presented, and the predictive ability in respect of a behavioral choice of both the classic and the newly developed algorithms is investigated and compared in an empirical study. Finally, the development of new tools for easily computing the IAT and SC-IAT scores is presented.

2.1 IAT D-score

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

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 $<$ 400 ms
*D*3	Mean $+$ 2*sd*	No
*D*4	Mean $+$ 600 ms	No
*D*5	Mean $+$ 2*sd*	Delete trials $<$ 400 ms
*D*6	Mean $+$ 600 ms	Delete trials $<$ 400 ms

Note: For all algorithms, trials with latency $>$ 10,000 ms 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 and B6, B7 are used for the computation.

The error correction strategies based on the built-in correction (*D1* and *D2*) refer to the IAT administration including feedback, as described in Section 1.2, for which error 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 increased 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 600 ms or twice the standard deviation). The other feature differentiating the *D-score* algorithms is the lower tail treatment, according to which fast trials (< 400 ms) are discarded or not.

Regardless of the specific guidelines for each algorithm, the core procedure for computing the *D-score* is the same. Firstly, the *D-score*s for the associative practice blocks (Eq. 2.1) and for the associative test blocks (Eq. 2.2) are computed:

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

and

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

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*:

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

The block orders in Equation 2.1 and Equation 2.2 is arbitrary and can be reversed. The resulting *D-score* has to be interpreted accordingly. Sticking with the Coke-Pepsi IAT structure depicted in Table 1.1, 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 SC-IAT D-score

Since blocks B1 and B3 (Table 1.2) are considered as pure practice blocks, they are not considered for 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 hence discarded from the computation. All responses with a latency faster than 350 ms 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 400 ms.

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* depend on that order.

Consider the Coke SC-IAT in Table 1.2. If the *D-score* is computed as the difference between Block B4 and Block 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

The aim of this study was to provide a fairer comparison of the predictive ability of the IAT and the SC-IAT in respect to a behavioral choice.

In Study 1, (Karpinski & Steinman, 2006) directly investigated and compared the predictive ability of a Soda IAT, that of a Coke SC-IAT, and that of a Pepsi SC-IAT, in respect to a soda, namely Coke vs Pepsi. As their results were suggesting, measures obtained from both the Soda IAT and the Pepsi SC-IAT played a role in predicting the actual soda choice, while the measure obtained from the Coke Sc-IAT did not contribute to the choice prediction. Drawing on this 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 & Steinman, 2006) provided interesting information on the functioning of implicit processes and the comparison between implicit measures, it also showed some shortcomings that might have undermined the validity of their results. Firstly, 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 affect by issues concerning both their administration and scoring procedures. The IAT and the SC-IAT differed in both the number of trials and the number of exemplars representing each category. The SC-IAT employed 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 (i.e., the stimulus was disappearing from the screen after 1,500 ms). Conversely, the IAT did not have such a constraint on the responses. Beyond making the task more difficult, the presence of a rtw produces a sense of urgency that it is otherwise missing (Karpinski & Steinman, 2006). Consistently with the SC-IAT administration procedure presented in Section 1.4, respondents were given feedback for each correct and incorrect response. No feedback, either for correct or incorrect responses, was given in the IAT. The procedures differed also on the labels used for representing the positive and negative evaluative dimensions (*Pleasant* vs *Unpleasant* for the IAT and *Good* vs *Bad* for the SC-IAT). Also the response keys used for sorting the stimuli hanged across measures. The IAT *D-score*s were computed according to the *D-score* procedure in Greenwald et al. (2003), despite Karpinski and Steinman (2006) fails to report the exact algorithm they have 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 capacity. It would be interesting to compare the predictive ability of the two implicit measures by using the same scoring procedure 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 method and by reducing the differences related to the administration procedures 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, response time window, 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 Milk chocolate SC-IAT, and one Dark 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 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 both implicit measures (i.e., 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 the possibility of withdrawing 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 of the category to which it was supposed 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. In all implicit procedures, evaluative attributes categories were composed by 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"). Object categories were labeled as *Dark* or *Milk*. 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 right side of the screen. Differently for the SC-IAT procedure described in Section 1.4, 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 a built-in correction or a response time window. Respondents did not receive any feedback on their performance, and they were asked to be as fast and accurate as they could in performing the tasks.

Additionally, respondents were explicitly asked to report their evaluation for Dark and Milk chocolate on two distinct items ("How much do you like Dark/Milk chocolate?") rated from "0 – Not at all" to "5 – Very much". The order of presentation of the implicit measures was counterbalanced across participants, and the explicit assessment was kept constant after the administration of implicit measures.

As a reward for their participation, respondents were offered with a free piece of Dark or Milk chocolate. 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.

Scoring. For the IAT, all *D-score*s 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*. Since the SC-IAT administration did not include a rtw and there are no guidelines concerning the upper response times treatment, no deletion of upper time responses was applied.

The scoring algorithms that have been introduced in this study result from different combinations of two main characteristics: The trials on which the standard deviation for the replacement of error responses was computed (i.e., only correct trials or all trials) and the quantity used for standardizing the difference in the response times between conditions (i.e., Cohen's pooled standard deviation or pooled trials standard deviation). The resulting eight combinations (identified by the letter "*m*") are reported in Table 2.2.

While the classical procedure for the SC-IAT includes a default lower tail treatment, the lower tail treatment for the IAT depends on the *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 350 *ms* 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 10,000 ms were discarded). Concerning the SC-IAT upper tail

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

Features	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$	$m7$	$m8$
Lower tail treatment	< 350 *ms*							
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

treatment, it might be argued that the deletion of the responses higher than 1,500 *ms* (i.e., the rtw cut-off) could 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 response time window 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 600 ms 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 (400 ms) 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 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.

Both classic and modified scoring for the IAT were computed so that positive scores indicated faster responses in associating Milk chocolate with positive attributes and Dark chocolate with negative attributes, and hence a likely preference for milk chocolate and/or a dislike for dark chocolate. Conversely, negative scores indicated faster responses in associating Dark chocolate with positive attributes and Milk chocolate with negative attributes, indicating a probable preference for Dark chocolate and/or a dislike for Milk chocolate.

For the SC-IATs, both the classic 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 a likely positive evaluation of the target chocolate. Conversely, negative scores indicated faster responses when the target chocolate was associated with negative attributes, and hence a probable negative evaluation towards the target chocolate.

2.3.2 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, 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 350 ms). 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 (first quartile = 2, third quartile = 5). The median for the explicit evaluation of Milk chocolate was 4 (first quartile = 3, third quartile = 4). No trials exceeding the threshold of 10,000 ms were found in the SC-IATs. Three trials exceeding the 10,000 ms threshold were found in the IAT, and they were eliminated. The lowest percentage of trials faster than both 400 ms (1.39%) and 350 ms (0.19 %) were found in the IAT. The two SC-IATs showed similar percentages of trials faster that 350 ms (1.00 % and 0.90 % in Milk SC-IAT and Dark SC-IAT, respectively), as well as of trials faster than 400 ms (4.40% and 4.32% in Dark SC-IAT and in Milk SC-IAT, respectively). All the implicit measures had the same overall percentage of correct responses (95%). The two SC-IATs showed similar average response latency (Dark SC-IAT $= 679.45 \pm 328.72$ ms, Milk SC-IAT $= 675.90 \pm 322.31$ ms). IAT average response time was 862.03 ± 496.50 ms. Descriptive statistics for 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 with explicit chocolate evaluations.

	Modified	*M (sd)*	*Min*	*Max*	r_{Milk}	r_{Dark}	Classic	*M (sd)*	*Min*	*Max*	r_{Milk}	r_{Dark}
IAT	*m 1*	$0.64\,(0.62)$	-1.91	1.72	0.40^{***}	-0.36^{***}	*D3*	$0.41\,(0.41)$	-1.29	1.25	0.42^{***}	-0.38^{***}
	m 2	$0.64\,(0.60)$	-1.86	1.69	0.40^{***}	-0.37^{***}	*D4*	$0.39\,(0.39)$	-1.26	1.27	0.41^{***}	-0.38^{***}
	m 3	$0.73\,(0.73)$	-2.25	2.84	0.39^{***}	-0.34^{***}	*D5*	$0.40\,(0.41)$	-1.29	1.29	0.42^{***}	-0.37^{***}
	m 4	$0.72\,(0.70)$	-2.12	2.59	0.39^{***}	-0.35^{***}	*D6*	$0.39\,(0.39)$	-1.26	1.32	0.41^{***}	-0.37^{***}
	m 5	$0.64\,(0.63)$	-2.29	1.72	0.40^{***}	-0.35^{***}						
	m 6	$0.64\,(0.60)$	-1.85	1.69	0.40^{***}	-0.36^{***}						
	m 7	$0.72\,(0.75)$	-2.34	2.85	0.39^{***}	-0.34^{***}						
	m 8	$0.72\,(0.71)$	-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^{*}
	m 2	$-0.06\,(0.34)$	-1.03	0.94	-0.23^{**}	0.17^{*}						
	m 3	$-0.06\,(0.36)$	-1.01	1.07	-0.22^{**}	0.17^{*}						
	m 4	$-0.06\,(0.35)$	-1.05	0.94	-0.22^{**}	0.17^{*}						
	m 5	$-0.06\,(0.36)$	-1.00	1.13	-0.20^{*}	0.16^{*}						
	m 6	$-0.06\,(0.35)$	-0.95	0.99	-0.20^{*}	0.16						
	m 7	$-0.06\,(0.36)$	-1.04	1.13	-0.19^{*}	0.16						
	m 8	$-0.06\,(0.35)$	-0.97	1.00	-0.20^{*}	0.16						
Milk SC-IAT	*m 1*	$0.16\,(0.39)$	-1.92	1.22	0.17^{*}	0.04	*D-Milk*	$0.15\,(0.33)$	-0.93	1.21	0.13	0.06
	m 2	$0.16\,(0.39)$	-1.93	1.13	0.17^{*}	0.04						
	m 3	$0.16\,(0.41)$	-1.92	1.5	0.17^{*}	0.04						
	m 4	$0.16\,(0.40)$	-1.94	1.38	0.17^{*}	0.04						
	m 5	$0.16\,(0.38)$	-1.39	1.23	0.15	0.05						
	m 6	$0.16\,(0.37)$	-1.40	1.14	0.15	0.05						
	m 7	$0.17\,(0.39)$	-1.39	1.51	0.16^{*}	0.05						
	m 8	$0.16\,(0.39)$	-1.40	1.39	0.16^{*}	0.05						

Regardless of the algorithm used, SC-IAT scores tended to have smaller effect sizes than IAT classic scores. IAT modified scores showed higher effect sizes than IAT classic scores, and this effect was particularly noticeable for the scores using Cohen's pooled standard deviation. Modified and classic SC-IAT scores were more consistent between each other.

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

To check the consistency between the scores, Pearson's correlations were computed between the classic and modified scores and the explicit chocolate evaluations.

Correlation coefficients between classic IAT *D-score*s ranged between $.99$ and 1.00 (all ps $< .001$). Correlations between classic IAT scores and Dark SC-IAT classic score were all $-.21$ (all ps $< .010$). No correlations were found between classic IAT *D-score*s and classic Milk SC-IAT score (correlations ranged between $-.04$ and $-.03$, all ps $> .050$). Classic *D-Dark* and *D-Milk* positively correlated between each other ($r = .15$, $p > .050$), but the correlation was not significant. Correlations between IAT modified scores ranged between $.97$ and $.99$ (all ps $< .001$). Their correlations with modified *D-Dark* ranged between $-.31$ and $-.28$ (all ps $< .001$). Modified *D-Milk* and modified IAT *D-score*s did not correlate between each other (correlation coefficients ranged between $-.01$ and $.01$, all ps $> .050$). Correlations between modified *D-Dark* ranged between $.98$ and 1.00 (all ps $< .001$). Correlations between modified *D-Milk* ranged between $.99$ and 1.00 (all ps $< .001$). Correlation between modified *D-Milk* and *D-Dark* showed the same direction as the correlation between classic 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. Both classic and modified scores were regressed on the behavioral chocolate choice, coded as 0 for Dark chocolate choice (DCC) and 1 for Milk chocolate choice (MCC). For the IAT, each score was regressed on the choice in a separate logistic regression. Since the choice is a dichotomous task in which Dark chocolate is contrasted with Milk chocolate, it is plausible that positive and negative attitudes towards both type of chocolates had played a role in determining the actual choice. The scores obtained from each of the SC-IATs convey a unique information regarding the positive or negative evaluation of only one type of chocolate, therefore lacking a part of information that might be crucial in determining the choice. It can also be argued that, since the two SC-IATs are two different experiments, it would not make sense to use the linear combination of their scores for predicting the actual choice. Grounding on these considerations, both single Dark SC-IAT and Milk SC-IAT scores, 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) were used as criteria for the investigation of the scores best accounting for the actual 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 total number of observed DCCs), and MCC accuracy (i.e., ratio between the number of MCCs correctly identified by the model and total number of observed MCCs) were computed. Results of the logistic regressions for predicting the chocolate choice are reported in Table 2.4.

Table 2.4: Choice prediction results: Single predictors.

	β *SE*	*Deviance* (*df* = 150)	*Nagelkerke's* R^2	General Accuracy	DCC Accuracy	MCC Accuracy
D-Score 3	2.23 *** [0.54]	188.22	0.18	0.64	0.63	0.66
D-Score 4	2.26 *** [0.55]	188.80	0.18	0.64	0.66	0.63
D-Score 5	2.18 *** [0.53]	188.92	0.18	0.63	0.63	0.63
D-Score 6	2.22 *** [0.55]	189.48	0.17	0.64	0.65	0.63
D-Dark	−0.70 *** [0.54]	208.77	0.01	0.53	0.65	0.41
D-Milk	0.35 *** [0.49]	209.99	0.01	0.53	0.78	0.26
IAT *m1*	1.35 *** [0.34]	191.07	0.16	0.64	0.65	0.63
IAT *m 2*	1.38 *** [0.35]	190.91	0.16	0.62	0.62	0.63
IAT *m 3*	1.07 *** [0.28]	192.62	0.15	0.63	0.65	0.62
IAT *m 4*	1.12 *** [0.29]	192.09	0.15	0.64	0.66	0.63
IAT *m 5*	1.35 *** [0.34]	190.95	0.16	0.64	0.65	0.63
IAT *m 6*	1.38 *** [0.35]	190.73	0.16	0.64	0.63	0.64
IAT *m 7*	1.07 [0.28]	192.51	0.15	0.63	0.65	0.62
IAT *m 8*	1.12 [0.29]	191.93	0.15	0.64	0.66	0.63
Dark *m1*	−0.73 [0.48]	208.07	0.02	0.53	0.62	0.44
Dark *m 2*	−0.72 [0.49]	208.24	0.02	0.53	0.62	0.42
Dark *m 3*	−0.70 [0.47]	208.18	0.02	0.53	0.62	0.44
Dark *m 4*	−0.69 [0.48]	208.33	0.02	0.52	0.62	0.41
Dark *m 5*	−0.62 [0.47]	208.65	0.02	0.52	0.63	0.40

Dark *m 6*	-0.63 [0.48]	208.72	0.02	0.52	0.63	0.40
Dark *m 7*	-0.60 [0.46]	208.74	0.02	0.51	0.63	0.38
Dark *m 8*	-0.60 [0.47]	208.80	0.01	0.51	0.63	0.38
Milk *m1*	0.33 [0.42]	209.86	0.01	0.53	0.77	0.26
Milk *m 2*	0.33 [0.43]	209.88	0.01	0.53	0.77	0.26
Milk *m 3*	0.32 [0.40]	209.83	0.01	0.52	0.76	0.26
Milk *m 4*	0.32 [0.41]	209.85	0.01	0.53	0.77	0.26
Milk *m 5*	0.31 [0.44]	209.99	0.00	0.50	0.75	0.23
Milk *m 6*	0.31 [0.44]	209.99	0.00	0.52	0.76	0.26
Milk *m 7*	0.30 [0.42]	209.95	0.00	0.50	0.75	0.23
Milk *m 8*	0.30 [0.42]	209.96	0.00	0.51	0.76	0.25

Note: ***: $p < .001$, **: $p < 0.10$, * $p < 0.05$. The Null Deviance for all the models is 210.48 ($df = 151$). βs are the log-odds for the probability of choosing Milk chocolate.

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

Classic and modified IAT scores tended to have similar values of both Nagelkerke's R^2 and accuracy of prediction Modified IAT scores tended to have slightly smaller values of Nagelkerke's R^2. All modified measures of 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 values than the classical ones. Regarding the accuracy of prediction, modified SC-IAT scores showed a slightly worse performance than the classic ones.

Results of choice prediction including the linear combination of SC-IAT scores is reported in Table 2.5.

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

	β_{Dark} *[SE]*	β_{Milk} *[SE]*	*Deviance*	*Nagelkerke's* R^2	General	DCC	MCC
Classic	$-0.77\,[0.55]$	$0.46\,[0.50]$	207.93	0.02	0.55	0.67	0.41
m 1	$-0.81\,[0.49]$	$0.44\,[0.43]$	206.99	0.03	0.56	0.66	0.45
m 2	$-0.80\,[0.50]$	$0.44\,[0.44]$	207.19	0.03	0.53	0.66	0.40
m 3	$-0.78\,[0.48]$	$0.43\,[0.41]$	207.07	0.03	0.54	0.66	0.41
m 4	$-0.77\,[0.49]$	$0.43\,[0.42]$	207.28	0.03	0.54	0.67	0.40
m 5	$-0.71\,[0.48]$	$0.44\,[0.45]$	207.70	0.02	0.55	0.68	0.40
m 6	$-0.72\,[0.49]$	$0.44\,[0.45]$	207.78	0.02	0.54	0.68	0.38
m 7	$-0.68\,[0.47]$	$0.42\,[0.43]$	207.77	0.02	0.55	0.68	0.40
m 8	$-0.69\,[0.48]$	$0.42\,[0.44]$	207.85	0.02	0.54	0.68	0.38

Note: ***: $p < .001$, **: $p < 0.10$, * $p < 0.05$. The Null Deviance for all the models is 210.48 ($df = 151$). βs are the log-odds for the probability of choosing Milk chocolate. General: General accuracy of prediction, DCC: Dark Chocolate Choice accuracy of prediction, MCC: Milk Chocolate Choice accuracy of prediction.

The linear combination of SC-IATs scores resulted in a better prediction of the chocolate choice than their singular scores. Nonetheless, IAT still outperformed them in the prediction of the choice. The coefficients of classic and modified *D-Dark* scores tended to be higher than the coefficients of classic and 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 classic scores and with the last four modified scores. General and DCC accuracy were similar across all the scores, while MCC accuracy showed a higher variability. Specifically, the linear combination of modified scores 6 and 8 showed the worst performance of all. The linear combination of modified scores 1 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, Dark and Milk chocolate explicit evaluations were included. IAT *D-score*s entered at the second step. *D-Dark* entered at the third step, and *D-Milk* entered at the fourth step. This procedure was followed for both modified and classic 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 modified and classic scores. It is reasonable to argue that the 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.

2.3.3 Discussion and Conclusions

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

The IAT outperformed both SC-IATs in the prediction of the behavioral choice. This result held both when using the separately the scores of the two SC-IATs both when their linear combination was considered. Taking these considerations together, it is possible to argue that the IAT has a higher predictive capacity than the SC-IAT.

However, the higher predictive capacity of the IAT might be due to the characteristics of the choice task itself. Since participants were presented with two different bowls of chocolate and were invited to take one piece of chocolate, 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 does, might hence best account for the actual chocolate preference, and result in a better accuracy of prediction. Conversely, measures dealing with only one of the components of the chocolate evaluation, like the SC-IAT does, might be disadvantaged. However, even when SC-IAT scores were considered concurrently, and hence both the chocolate evaluations were included, the general accuracy of prediction was similar to the one obtained when the single measures were considered in the choice prediction. Nonetheless, it resulted in a slightly better accuracy of prediction for the Milk chocolate, especially when compared to the performance of the Milk SC-IAT scores alone. This result is supporting the claim for which a measure including the comparative evaluation between two contrasted objects, instead of their absolute evaluation, results in better prediction of the choice between alternative options.

Finally, the positive evaluation of one type of chocolate does not imply the negative evaluation of the other. 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 in cases in which a clearly contrasted category is not identifiable, like in the self-esteem case. In this case, it is possible to argue that the SC-IAT might perform better than the IAT. Moreover, the SC-IAT might also perform better than the IAT in predicting a behavioral choice when the choice is not strictly dichotomous like it was in this study. For instance, respondents might be left free to choose between one type of chocolate, both types, or none of them, or even between different types of candy bars, including Dark and Milk chocolate bars.

Both in the study by Karpinski and Steinman (2006) and in the present one, the implicit measures predictive validity 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. An example can be 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 response time window at 1,500 ms. In the present study, no response time window was used. We could have applied an a posterior threshold for upper tail responses as if a response time window had been used in the administration of the implicit measure itself. However, we decided not to do so because using a response time window affects respondents' performance. Indeed, Karpinski and Steinman (2006) observed that the response time window influenced the speed of the participants in the sense of producing a sense of urge for giving the response that was missing when the response time window 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 `implicitMeasures`

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

Table 2.6 offers an overview of the existing `R` packages and their functions for the computation of the IAT *D-score*.

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

The packages for computing the IAT *D-score* present some shortcomings. For instance, packages **IATanalytics** and **IATscore** require for a specific arrangement of the data set columns to compute the *D-score*. Moreover, both packages compute the *D-score* for one respondent at a time. **IAT** package asks for a counter-intuitive coding of accuracy responses (i.e., 0 for correct responses, 1 for error responses). **IATScores** package appears to be the most complete package, but it does not allow for scoring other implicit measures. Plotting functions are available in only two packages (**IAT** and **IATScores**), but in both cases the functions are not meant for plotting the *D-score*s.

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 the exact label of the *D-score* algorithm they decided to use, raising replicability issues (Ellithorpe et al., 2015). Replicability issues might also rise from the many steps needed to clean and prepare the starting data set for computing the *D-score* (Ellithorpe et al., 2015). Only two packages out of four (**IAT** and **IATScores**) provide functions for computing the different *D-score* algorithms.

The aim with which **implicitMeasures** was developed is to provide an `R` package that clearly disambiguate the algorithms and all the undertaken steps for computing the *D-score*, making the *D-score* computation easier and mistakes free.

The source code of **implicitMeasures** is available on GitHub. The package is downloadable from CRAN.

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

implicitMeaures provides an easy way to compute the algorithms for the both the IAT and the SC-IAT in an automated way, hence preventing computational mistakes occurring during data preparation and *D-score* computation. 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. The possibility to compute all the 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 are

Table 2.7: **implicitMeasures** contents and functions.

Function	Description
`clean_iat()`	Prepare and clean IAT data
`clean_sciat()`	Prepare and clean SC-IAT data
`computeD()`	Compute IAT D-score
`Dsciat()`	Compute D for the SC-IAT
`descript_d()`	Descriptive table of the IAT or SC-IAT *D-score*
`d_distr()`	Plot IAT or SC-IAT scores (distribution)
`d_plot()`	Plot either IAT or SC-IAT scores (points)
`IATrel()`	IAT reliability
`multi_dsciat()`	Plot SC-IATs scores
`multi_dscore()`	Compute and plot multiple IAT *D-score*s
`raw_data()`	Data set with one IAT and two SC-IATs

affecting the results. By preventing computational mistakes, providing a clear reference on the specific *D-score* that has been computed, and allowing for the comparison among *D-score* algorithms, **implicitMeasures** can hence improve the replicability of the results.

All the objects created with **implicitMeasures** functions are exportable in external files. For example, the data frame obtained from the `clean_iat()` function can be easily exported in a CSV file, which can then be uploaded to DscoreApp (Epifania, Anselmi, & Robusto, 2019a, see Section 2.4.2). .

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

2.4.2 DscoreApp

Beyond the `R` packages illustrated in the previous section, also other options are available for the computation of the IAT *D-score*, namely Iniqusit scripts and SPSS syntaxes.

Inquisit scripts are the most straightforward way for obtaining the *D-score*, since they compute it right after the IAT administration procedure, they store the result along with other information on participants' performance (e.g., response time for each IAT trial, correct and incorrect responses), and they do not require any programming skills. Nonetheless, these scripts work only when associated with the Inquisit administration procedure, they can compute just one of the available *D-score* algorithms, and they do not provide functions for visually inspect 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 as well as a certain degree of expertise with SPSS language. They do provide the chance to compute different *D-score* algorithms. No functions for directly plotting the results are provided.

`R` packages provides the open source alternative to both Inquisit and SPSS syntaxes, but, as it has been highlighted in the previous section, their use is not always straightforward, they require quite advanced programming skills, and their functions are often not well defined and understandable. Even though `implicitMeasures` addresses the majority of the shortcomings of the `R` packages available for the computation of the *D-score*, it still requires a certain familiarity with programming to be used efficiently.

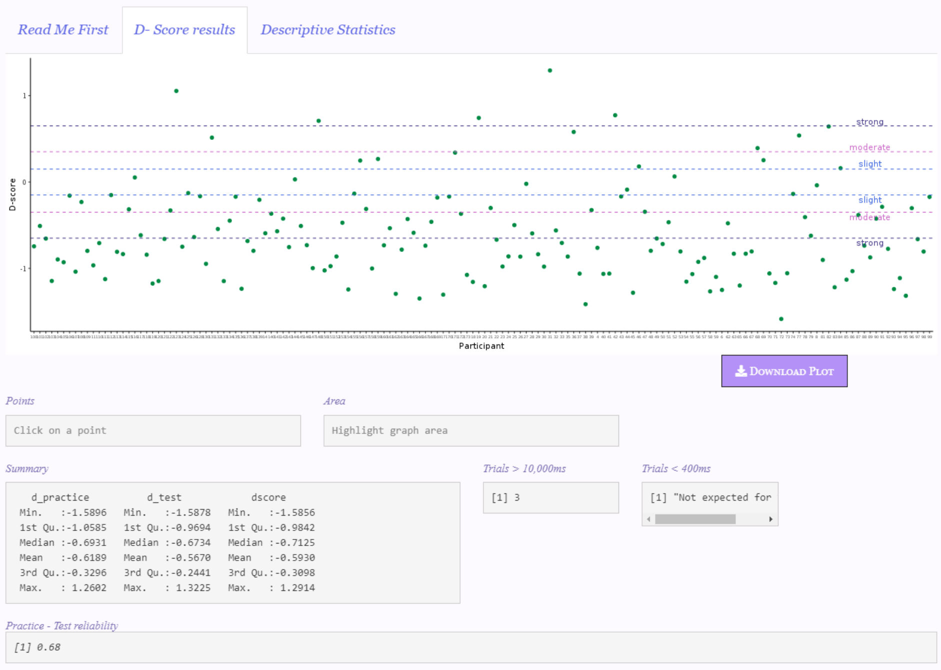
DscoreApp was developed in `R` by means of `shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2018) and `shinyjs` (Attali, 2018) packages with the aim of providing an open source tool able to make the *D-score* computation easier fore researchers who commonly employ the IAT but have little or no programming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

The app is organized in different panels ("Input", "Read Me First", "D-score results", and "Descriptive statistics"), and it provides users with a toy data set that can be used to familiarize oneself with DscoreApp functions. The setting options and functions in the "Input" panel and 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. It includes a downloadable template suggested for using the app (i.e., **Download template** button), even though it is not necessary to use it. Indeed, DscoreApp is designed to work as long as the uploaded data set is in a CSV format, using comma as the column separator, and includes the following variables: `participant` (i.e., participants' IDs), `latency` (i.e., response latencies in milliseconds), `correct` (response accuracy, either 0 for error responses or 1 for correct responses), `block` (i.e., labels identifying the four associative blocks of the IAT). This panel also contains information on the downloadable file that can be retrieved after the *D-score* computation.

Users can upload their own data set (i.e., by using the **Browse** function), or use the toy data set included in the app(i.e., by checking the `Race IAT dataset` checkbox) in the "Input panel" Once the data set is read, the app automatically populates the drop-down menus for choosing the labels of the four associative blocks, and the **Prepare data** button is activated. When data are ready for the *D-score* computation, the "Data are ready" message appears next to the **Prepare data** button, and all the options for its computation and graphical display become active. Once a *D-score* algorithm is chosen from the drop-down menu "Select your D", the **Calculate & Update** button is activated. 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 ($<$ 300 ms) exceed 10% of the total responses (Greenwald et al., 2003). When these options are selected, participants exceeding the thresholds (if any) will not be 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 "Input panel". The `Summary` box reports the descriptive statistics 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 trials discarded because of fast response times. This box is populated only when



Points

Click on a point

Area

Highlight graph area

Summary

```
  d_practice          d_test            dscore
 Min.   :-1.5896   Min.   :-1.5878   Min.   :-1.5856
 1st Qu.:-1.0585   1st Qu.:-0.9694   1st Qu.:-0.9842
 Median :-0.6931   Median :-0.6734   Median :-0.7125
 Mean   :-0.6189   Mean   :-0.5670   Mean   :-0.5930
 3rd Qu.:-0.3296   3rd Qu.:-0.2441   3rd Qu.:-0.3098
 Max.   : 1.2602   Max.   : 1.3225   Max.   : 1.2914
```

Trials > 10,000ms

```
[1] 3
```

Trials < 400ms

```
[1] "Not expected for
```

Practice - Test reliability

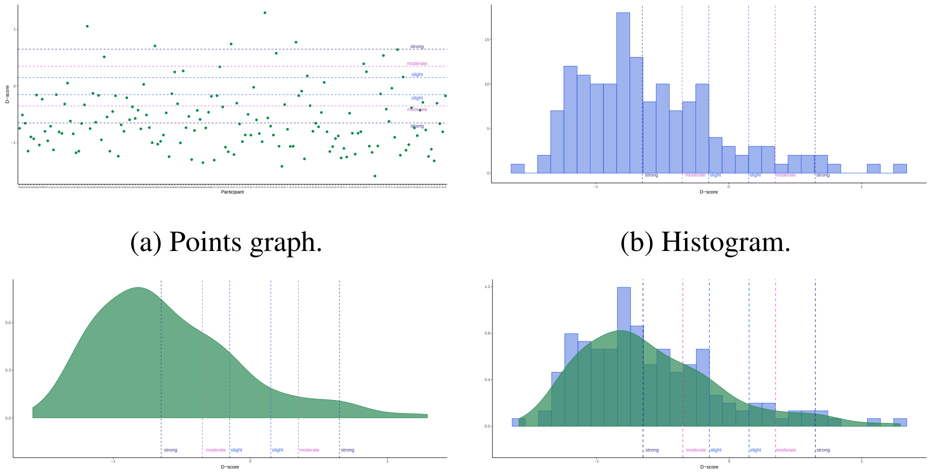
```
[1] 0.68
```

Figure 2.1: DscoreApp results panel.

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).

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

Graphical representation is a convenient way to identify extreme scores or particular response pattern. 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 `Point` box (see Figure 2.1). By highlighting an area of the graph, the IDs of participants included in the area, along with their *D-score*s appears in the `Area` box (see Figure 2.1).



(a) Points graph.

(b) Histogram.

(c) Density.

(d) Histogram and Density.

Figure 2.2: Available graph representation

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

Chapter 3

Formal modeling

In this chapter, a brief overview of the models proposed for modeling IAT data is provided, and their advantages and drawbacks are highlighted. Some of these models are aimed at investigating the cognitive processes that are playing a role in the IAT by exploiting all the information that can be gathered from the accuracy responses (Section 3.1), while others are concurrently considering the accuracy and response times for delving deeper on the implicit associations driving the performance (Section 3.2). One common drawback of these models is that they are not able to provide information at the level of the individual stimulus.

The advantages of Rasch modeling of the IAT data by means of the Many Facet Rasch Modeling (MFRM) are illustrated in Section 3.3. Although this approach can provide stimulus–specific information, it also comes with drawbacks that cannot be ignored.

3.1 Multionomial Models

3.1.1 Quad Model.

The Quad model (Conrey et al., 2005) is a multinomial model introduced for distinguishing the contribution of automatic and controlled process in driving the respondents' performance in the IAT. Specifically, Quad model is aimed at highlighting the contribution of four qualitatively different processes defined by different levels of automaticity and/or controllability. It is completely based on accuracy responses, and it exploits the logic of response compatibility. Response compatibility is the assumption on which the IAT rests, namely, that responses are more accurate and faster in the condition compatible with respondent's automatic association. The activation, or lack of thereof, of the four processes contributes to the overt response in the IAT. The observed probability of a correct response to a particular category of stimuli is the starting point for estimating the four parameters. The sum of the probabilities associated with a response is the total probability of that response.

The likelihood that an automatic association is activated by a category of stimuli is captured by the *association activation* parameter (AC), which 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, the more likely it will be activated by the triggering stimulus.

The probability of giving the correct response also depends on the ability of identifying the actual correct response, and the *discriminability* parameter (D) expresses the likelihood that the correct response (i.e., the correct category to which the triggering stimulus belongs) is identified.

If a negative automatic association is activated by the stimulus (e.g., racial prejudice), the respondent might try to fake the response in a socially desirable way. This process is captured by the *overcoming bias* parameter (OB), which expresses the likelihood that an activated bias is overcome in favor of a deliberate, and probably more desirable, response.

Parameters D and OB are the parameters that are mostly influenced by motivation and that mostly depend on the availability of cognitive resources to be effective.

Finally, the Quad model takes into account the chance that, in absence of any other of the processes, the responses are influenced by any response bias, which might also be tendency to respond with the left response key. The *guessing* parameters (G) defines the likelihood that this response bias is activated and drives the responses.

Despite this model results in a distinction between automatic and controlled processes taking place during the IAT, and can hence provide a better understanding of the IAT functioning, it presents some drawbacks. Since it is based solely on accuracy responses, it can present at least two issues. Firstly, the IAT is designed to be an extremely easy task, hence the proportion of correct responses is extremely high. Secondly, this model is neglecting all the information that can be gathered from the responses times, and that can potentially lead to further and more detailed results. Finally, all the parameters are at the sample level.

3.1.2 ReAL Model

The ReAL model (Meissner & Rothermund, 2013) is a multinomial processing tree model aimed at disentangling the different processes involved in the IAT. The assumption at the base of this model is that part of the respondents' performance is not ascribable to their automatic association, but it is due to strategies undertaken during the administration. Thus, this model should be able to mathematically distinguish between actual attitude associations and recoding strategies.

As the Quad model, the ReAL model is based on the accuracy responses. Three processes are assumed to drive the correct and incorrect response pattern in the IAT, according to the IAT associative condition. Specifically, in the condition compatible with the automatic association of the respondent, all three processes are supposed to play a role, while in the incompatible condition only two of them are driving the response pattern.

By assuming that attitude objects can activate evaluative associations, a Re-coding process (i.e., *Re*) can intervene in the compatible condition. This means that the categorization tasks is reduced from four categories to two categories by exploiting the evaluative characteristics of the attitude objects. In the compatible condition, the object stimuli can be sorted by using their attribute characteristics (i.e., being evaluated as *Good* or *Bad*). The Re-coding process can facilitate the task in the compatible condition, while it can interfere with it in the opposite one. If the Re-coding processes is not activated, then the Label-based (i.e., *L*) process drives the response. This process represents the controlled search for the label representing the category to which the stimulus belongs. When neither the Re-coding process nor the Label-based one are involved, the Evaluative associations (i.e., *A*) of the object categories can actually drive the response.

Evaluative associations and label based processes take place in both IAT associative conditions. Recoding takes place only in the condition that is consistent with the respondent's automatic activated association.

3.2 Time and Accuracy models

3.2.1 Diffusion Model

The first application of the Diffusion Model (DM; Ratcliff, 1978) to the IAT data can be found in Klauer et al. (2007). As the Quad and the ReAL models, also the DM is aimed at disentangling the contribution of different process contributing to performance at the IAT. Additionally, since DM maps accuracy and response times on the same metric, it can exploit the information deriving from both type of responses.

The assumption on which the DM rests is that decisions in two-choice tasks like the IAT are based on processes of information accumulation during time. Once the amount of information reaches the thresholds for determining the responses, the response associated to that threshold is given.

According to DM, there are four main components that intervene in information processing. The first one is the speed with which the information is accumulated. This components is expressed by the *drift rate* (i.e., parameter v), which quantifies the direction and speed with which the information is accumulated. It determines both the respondent's accuracy and time performance.

The thresholds defining the amount of information needed for giving a response are another component of the DM. Each threshold is associated with a response, namely the correct (stimulus correctly sorted in the category) and incorrect (stimulus incorrectly sorted in the incorrect category) one. This component si expressed by the *separation* parameter (i.e., parameter a), which quantifies the amount of evidence in favor of one versus the other response. Clearly, the amount fo information that needs to be accumulated depends on the starting point from which the accumulation process starts, that is, the *starting point* of information accumulation. This parameter informs about the response bias towards one of the two response thresholds.

Finally, there is also a non-decision component involved in the response process. This component refers to the time used for actions that are not directly implicated in the response process, like the encoding of the stimuli or the response execution itself. Parameter t_0 represents the contribution of the non-decision component to the response processes.

According to the DM, attitudes enter the IAT, and hence guide the responses, through the object stimuli. Specifically, the automatic association positively influences the *drift-rate* in the compatible condition (i.e., faster and more accurate responses), while it negatively influences the *drift-rate* in the incompatible condition (i.e., slower and less accurate responses). Consistently with the Recoding strategy illustrated in the ReAL model, the faster and more accurate responses in the compatible condition result from the use of the positive/negative valence of the object stimuli for their categorization, hence avoiding the task-switch cost from attribute to target.

Since the DM provides the distribution of both correct and incorrect responses, it results in a detailed information on the IAT functioning and on the respondents' performance. However, for doing so, it needs a high proportion of incorrect responses, which is quite difficult to observe in an IAT. One solution to make the respondents commit more mistakes is to add more trials to the each associative condition. Moreover, DM is separately applied to each entire critical block, without distinguishing between the different stimuli categories. This also implies that parameters values estimated from different block cannot be compared.

3.2.2 Discrimination-Association Model

The Discrimination-Association Model (DAM; Stefanutti et al., 2013) has been specifically developed for the analysis of IAT data, with the aim of separating automatic from controlled processes while accounting for both accuracy and time responses.

DAM decomposes the IAT effect into three processes, namely stimuli discrimination, automatic association, and termination criteria. Stimuli discrimination informs about the functioning of the stimuli and, specifically, whether the stimuli can be easily recognized and categorized into the correct category. However, this information is not at the level of the individual stimulus but at that of the stimuli category. The process that mostly regards individual differences is the automatic association one, which provides information about the association between the evaluative dimension and the object categories. This information allows for a better understanding the meaning of the IAT measures. Consider again the Coke-Pepsi IAT in Section 1.2. A positive *D-score* on this IAT might both indicate faster responses in associating Coke/Good, Pepsi/Bad, or both. In any case, a positive score is obtained, but with different meanings. If the Coke/Good association are driving the responses, it means that it is more the preference for Coke to guide the performance. Otherwise, if the associations Pepsi/Bad are driving the responses, it is more the dislike for Pepsi, than the preference for Coke, to determine the performance. However, the *D-score* cannot distinguish between the two distinct automatic associations, while DAM can, providing a better understanding of the IAT measures. Finally, termination criteria represents the amount of information that must be accumulated before a response is given. It can be interpreted as either task difficulty or individual cautiousness in giving the responses.

DAM provides useful information on the IAT functioning, and it also overcomes some of the issues of DM. For example, since the DAM assumes that correct and incorrect responses are deriving from different processes, reliable parameters estimates can be obtained even when no error responses are present. Besides, DAM is applied on the entire set of IAT blocks.

However, also DAM presents some shortcomings. Firstly, the information on stimuli functioning regards only the stimuli categories and not the individual stimulus. Secondly, since the termination criteria can be interpreted as either a property of the task (i.e., difficulty) or an individual characteristic (i.e., respondent's cautiousness), it is confounding the distinct contribution of the task and that of the respondent in determining the responses. Finally, DAM is based on Poisson processes, and it is actually an extension of the Poisson race model. The average user of the IAT tends not to be familiar with that kind of models, and might find the interpretation of the parameters, or the model itself, utterly complicated, without noticing the clear advantages deriving from this approach.

3.3 Rasch Modeling

By applying the Rasch model (Rasch, 1960, See Chapter 4.1) to IAT data it is possible to disentangle both respondents and stimuli contribution to each response process. The detailed and fine-grained information at both levels can be used to shed a new light on the meaning of the IAT measure. The information at the stimuli level can be used to test whether IAT stimuli are recognizable as prototypical exemplars of the category they are supposed to represent. Moreover, this information allows for investigating the contribution of each stimulus to the IAT effect, which can in turn help in understanding the meaning of the IAT measure. The information at the respondents' level might help in clarifying whether there are any artifacts in the IAT procedure that can affect respondents' performance, and hence the resulting IAT effect.

The Many Facet Rasch Model (MFRM; Linacre, 1960) has been successfully applied to the discretized IAT time responses, providing interesting insights on the functioning of the IAT stimuli (Anselmi et al., 2011; Anselmi, Vianello, Voci, & Robusto, 2013; Anselmi, Vianello, & Robusto, 2013). The information obtained on the stimuli was used for disentangling the automatic associations that mostly contribute to the IAT effect. In the Rasch model, the likelihood of a correct responses depends on both the respondent's ability θ and the stimulus difficulty b. The MFRM is an extension of the classic Rasch model in which other sources of systematic variability (i.e., *facets*) are accounted for the correct response.

The IAT associative condition can be considered as a facet of the MFRM, and, as such, it allows for obtaining condition–specific stimuli parameters. For instance, by comparing the condition–specific stimuli parameters, (Anselmi et al., 2011) found a positive primacy effect, for which the IAT effect was mostly driven by positive stimuli. By drawing on this result, the authors suggested that the implicit preference for Caucasian people over African people that is often observed in Caucasian respondents could be expression of in-group favoritism rather than out-group derogation.

Despite Rasch modeling of the IAT data provides interesting insights on the IAT functioning, it also presents some shortcomings. Firstly, by discretizing the response times, a part of the information that can be retrieved is lost, and the variability of the responses is reduced. Secondly, by applying the MFRM to the IAT data, its fully-crossed structure is not accounted for, hence the non-independence of the observations was not accounted for. Thus, the estimates obtained through the application of this model might result biased by the influence of unwanted and uncontrolled error variance. Finally, in the application made by Anselmi et al. (2011), the focus was on the functioning of the stimuli between the two associative conditions. To obtain the condition–specific estimates, it was assumed that the difficulty of the two conditions was assumed to be the same for all respondents. Consequently, it was not possible to investigate the respondents' individual differences in performing the IAT.

Chapter 4

Rasch model model and log-normal model

In this Chapter, Rasch model and log-normal models are briefly outlined. The advantages of using Linear Mixed Effect Models on IAT data are then discussed, as well as how to obtain the Rasch model and the log-normal model parameters estimates from the (Generalized) Linear Model, also considering the extension to multidimensional data.

4.1 Rasch Model

According to the Rasch model, the likelihood of a respondent i to endorse the correct response on a item j can be expressed as the function of the respondents ability θ and the item difficulty b. The correct response will be endorsed when the respondent ability is higher than the item difficulty. Vice-versa, an incorrect response will be endorsed when the respondent ability has a lower ability than the difficulty of the item. When the respondent's ability is equal to the stimulus ability, the respondent probability of endorsing the correct response will be equal to his/her probability of endorsing the incorrect response.

The Rasch model can be formally expressed as follow:

$$P(x_{ij} = 1|\theta, b) = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)}, \tag{4.1}$$

where x_{ij} is the accuracy response of respondent i on item j.

In the IAT case, the correct response to each stimulus is given by the correct categorization of the stimulus in the category to which it belongs. Respondent's ability can hence be considered as the ability to correctly recognize the correct category to which the stimulus belongs and assign it to that category. Stimulus difficulty can be considered as the stimulus characteristics that are making it more or less recognizable as a prototypical exemplar of the category it is supposed to represent.

The higher the value of θ_i, the higher the respondent ability to sort the stimuli into the correct categories, and hence the higher the proportion of responses correctly identified. The higher the value of b, the more difficult a stimulus is, meaning that it is not easily identified as a prototypical exemplars of the category to which it belongs, and it hence gets a higher proportion of incorrect responses.

4.2 Log-normal Model

The discretization of the response times needed for the application of the MFRM (see Chapter 3.3) can be avoided by considering the normal distribution of the log-transformed response times. By doing so, it is possible to consider the response times in their continuous nature while obtaining a parametrization similar to Rasch parametrization by means of the log-normal model (van der Linden, 2006).

According to the log-normal model, the log-time response of respondent i on item j can be expressed as a function of respondent's speed τ and stimulus time intensity δ. Time intensity can be defined as the time each stimulus requires for getting a response, and it is strongly dependent on the characteristics of the stimulus itself (van der Linden, 2006).

The log-normal model can be formally expressed as follow:

$$f(t_{ij}|\tau_i, \delta_j) = \frac{1}{t_{ij}\sqrt{2\pi}} exp \left\{ -\frac{1}{2}[ln\ t_{ij} - (\delta_j - \tau_i)]^2 \right\}, \tag{4.2}$$

where t_{ij} is the response time of respondent i on stimulus j. The person parameter τ describes the speed of the respondent; The larger the value of τ, the smaller the amount of time the person needs to respond. Stimulus parameter δ is interpreted as the time a stimulus demands for getting a response (i.e., *time intensity*); the larger the value of δ, the more time is required by the stimulus to get a response.

The relation between the stimulus parameter δ and the respondent parameter τ, expressing the mean of the distribution of Equation (4.2), resembles the relation between the respondent parameter θ and stimulus parameter b in the Rasch model (4.1).

4.3 Linear Mixed Effects Models

A Rasch approach to the analysis of the IAT data proven to be an efficient way to obtain stimuli parameters estimates leading to a deeper comprehension of the meausure itself. The use of a model considering the normal distribution of the log-time responses allows for considering the response time in their continuous nature and hence to overcome the issue of the loss of information due to the discretization for the application of the MFRM. Nonetheless, there is still something missing, that is, the cross-classified structure of the data is still not taken into account.

To deal with the cross-classified structure characterizing both the IAT and the SC-IAT, the multilevel structure of the data must be considered as reflecting the random variability of the population from which the different levels are drawn (Barr et al., 2013). This can be done by considering the random variation of each level and by specifying it as a random effect (Doran, Bates, Bliese, & Dowling, 2007).

4.3.1 (Generalized) Linear Model and Rasch Model

A generalized linear model is a linear model in which the linear predictor for the i-th response is related to the expected value of the response, μ_i by an invertible link function (g, McCullagh & Nelder, 1989). The combination of the linear predictors can be written as $\eta_i = x_i\beta$, where eta_i is the the combination of linear predictors and x_i is the i-ith row of the *observations* × *predictors* model matrix ***X***, derived from the model and it covariates.

The relationship between the linear predictors and the expected value of the response is as follows:

$$x_i\beta = \eta_i = g(\mu_i). \tag{4.3}$$

The *logit* is the natural link function for a binomial response (McCullagh & Nelder, 1989), defined as the logarithm of the odds (i.e., the ratio between the probability of a correct response and the probability of an incorrect response), for which:

$$\eta_i = g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right), \tag{4.4}$$

with inverse link:

$$\mu_i = g^{-1}(\eta_i) = \frac{exp(\eta_i)}{1+exp(\eta_i)}. \tag{4.5}$$

The structure of inverse of the *logit* link function in Equation **??** clearly allows for seeing the relationship with the Rasch model as expressed by the Model in Equation 4.1.

4.3.2 Extending to Multidimensional data

To extend the generalized linear model to multilevel data, and hence to move from a GLM to a Generalized Linear Mixed Effect Model (GLMM), the linear combination of predictors η must include both fixed effects β and random effects d, such that:

$$\eta = X\beta + Zd, \tag{4.6}$$

where X is the *observations* × *predictors* model matrix, whereas Z is the *observations* × *grouping factor*. Each component of the random effects d is associated with a level of the grouping factor.

At this point, the random variability of different grouping factors, being the type of items or the conditions of presentation, are included in the model matrix and are hence accounted for, and it is possible to obtain the Rasch model parameters by using the estimates of the fixed effects and the marginal modes of the random effects, or Best Linear Unbiased Predictors (BLUP). BLUPs are expressing the adjustment of each level of the random effects to the fixed effects. Formally, BLUPs are not parameters of the model themselves, and they are obtained after the estimation of the random effect parameters, that is, the expected variance of the specific level of the random effect.

Chapter 5

Models specification

The (Generalized) Linear Mixed Effect models random structures are specified in this chapter. Form each of the illustrated random structure, different parameters for the Rasch and log-normal models.

Since the IAT has a specific data structure, a Maximal Model was introduced. Nonetheless, the random structure of the Maximal Model is know to be too complex to fit on empirical data sets. Consequently, models with simpler random structures were specified starting by the random structure of the maximal model. A final overview of the Rasch and log-normal model that can be extracted from each of the random structures is presented.

5.1 Maximal Model

All the sources of variability in the IAT can be taken into account by specifying them in a model with the appropriate random structure, in which the error variance is decomposed into each of the sources of variation (Maximal Model, MM; Barr et al., 2013). The expected response y for the observation $i = 1, \ldots, n$ for participant $j = 1, \ldots, J$ on stimulus $k = 1, \ldots, K$ in condition $l = 1, \ldots, L$ can be expressed by the following model:

$$y_i = \alpha + \beta_i l_i + \alpha_{j[i]} + \alpha_{k[i]} + \alpha_{j[i]}\alpha_{k[i]} + \beta_{j[i]} l_i + \beta_{k[i]} l_i + \epsilon_i, \tag{5.1}$$

with:

$$(\alpha_j, \beta_j) \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_{\alpha_j} & \sigma^2_{\alpha_j,\beta_j} \\ \sigma^2_{\alpha_j,\beta_j} & \sigma^2_{\beta_j} \end{pmatrix}\right), \tag{5.2}$$

$$(\alpha_k, \beta_k) \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_{\alpha_k} & \sigma^2_{\alpha_k,\beta_k} \\ \sigma^2_{\alpha_k,\beta_k} & \sigma^2_{\beta_k} \end{pmatrix}\right), \tag{5.3}$$

$$(\alpha_j \alpha_k) \sim \mathcal{N}\left(0, \sigma^2_{\alpha_{j,k}}\right), \tag{5.4}$$

$$\epsilon \sim N(0, \sigma^2). \tag{5.5}$$

The model Fixed effects are expressed by the fixed intercept α and the fixed slope β. The fixed intercept α represents the expected average in the of the IAT condition and the fixed slope β expresses the deviation from the first IAT condition. By summing up together, it is possible to obtain the estimate of the expected average in the second IAT condition.

The between–respondents variability (i.e., difference in the response due to respondents individual differences) are accounted for by specifying their random intercepts $\alpha_{j[i]}$. Likewise, the between–stimuli variability is taken into account by specifying the stimuli random intercepts, represented by the term $\alpha_{k[i]}$. The variability due to the interaction between the stimuli and the respondents (i.e., respondents' individual reactions to each stimulus) is accounted for by the specification of the interaction effect between the stimuli and the respondents random intercepts (i.e., $\alpha_{j[i]}\alpha_{k[i]}$).

The between–conditions variability is accounted for by specifying the respondents and stimuli random slopes in the associative conditions. The random slopes express the impairment/facilitation effect of the IAT condition on the respondents' performance or on the stimuli categorization.

The MM, beyond being able to control for all the possible sources of random variation, is the one that allows for obtaining the higher amount of information. This model results in overall respondents' and stimuli parameters, as well as condition–specific respondents' and stimuli parameters.

The comparison between respondents' overall abilities and each of their two condition-specific parameters can inform about the bias in respondents' performance that is due to the IAT associative condition. Moreover, by computing the difference between the respondents' condition–specific parameters, it is possible to obtain a measure of this bias.

Likewise, by comparing each stimulus condition–specific parameters with their overall parameter allows for investigating whether the stimuli categorization is somehow affected by the IAT associative condition. By computing the difference between the condition–specific stimuli parameters, a measure of the bias due to the IAT condition can be obtained.

The MM is indeed providing numerous important information. Nonetheless, if the final aim is to obtain the Rasch and log-normal model parameters estimates from the IAT data, a model of such a complexity might not be needed. Moreover, a model of such a complexity is at risk of over-fitting or convergence failure (Bates, Kliegl, Vasishth, & Baayen, 2015), especially if the observed data do not have enough variability.

Grounding on these considerations, models with simpler random structure were specified, starting from the random structure of the MM. Although interesting, it is not necessary to obtain both overall and condition–specific parameters for both the stimuli and the respondents. The model can hence be simplified by dropping the estimation of these parameters. Furthermore, the estimation of the interaction effect between the respondents' and the stimuli random intercepts can be dropped as well. Indeed, the estimation of this parameter requires a high variability in the data, and, if there is not, it brings to convergence failure. The estimation of this parameter should be added to the model only in those cases in which the error variance is still high after the estimation of all the other parameters (Westfall et al., 2014). For estimating the Rasch and the log-normal models parameters, either the respondents' or the stimuli parameters have to be centered at 0, otherwise the model is not identified (?, ?; Gelman & Hill, 2007).

5.2 Models random structure

Three models for the accuracy responses (Rasch models) and three models for the log-time responses were specified. The random structures of these models were the same, besides the assumption on the error term ϵ. This parameter was assumed to follow a logistic distribution for the accuracy models ($\epsilon \sim \mathcal{L}(0, \sigma^2)$) and a normal distribution for the log-time models ($\epsilon \sim \mathcal{N}(0, \sigma^2)$).

Since the random structure is the same, and the only things that are changing are the assumptions on the error term and the dependent variable (either the accuracy or the log-time responses), only the equations for the specification of the accuracy models for the estimation of the Rasch models parameters are reported. The parameters of the log-normal model obtained through the application of the log-time models are commented as well. For both the accuracy and the log-time models, the fixed intercept α is set at 0, so that the estimates β of the fixed effect of the associative condition can be interpreted as the expected proportion of correct responses in each condition (for the accuracy models) and as the expected average response time in each condition (for the log-time models)

In Model 1, the within–respondents between–conditions variability and the between–stimuli variability are accounted for by specifying respondents' random slopes in conditions and stimuli random intercepts, respectively:

$$y_i = logit^{-1}(\alpha + \beta_i l_i + \alpha_{k[i]} + \beta_{j[i]} l_i + \epsilon_i), \tag{5.6}$$

with:

$$\beta_j \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_{\beta_{jl_1}} & \sigma^2_{\beta_{jl_1}, \beta_{jl_2}} \\ \sigma^2_{\beta_{jl_1}, \beta_{jl_2}} & \sigma^2_{\beta_{jl_2}} \end{pmatrix}\right), \tag{5.7}$$

$$\alpha_k \sim \mathcal{N}(0, \alpha^2_k). \tag{5.8}$$

The random structure of Model 1 results in condition–specific respondents' estimates and overall stimuli estimates. The condition–specific ability parameters of the Rasch model τ_{jl} are expressing the respondents' ability to perform the categorization task according to the associative condition. The higher the value of θ_{jl}, the higher the ability of the respondent to perform the categorization task in the specific condition, and hence the higher the proportion of stimuli correctly categorized in that condition. Similarly, the condition–specific respondents' speed parameters τ_{jl} of the log-normal model are expressing the speed with which the respondents' are performing the categorization task in the specific associative condition. The lower the valued of τ_{jl}, the higher the speed of the respondent, and hence the the lower the time needed for categorizing the stimuli. For both the Rasch and the log-normal models, the condition–specific respondents' parameters allow for testing whether there is an effect of the IAT associative condition on the respondents' accuracy or time performance, respectively. Moreover, by computing the difference between the respondents' condition–specific parameters it is possible to obtain a measure of the bias in the performance due to the associative condition. The overall stimuli estimates, either easiness b_k or time intensity δ_k, can inform about each stimulus representativeness of its alleged category. The higher the easiness parameter b, the easier the stimulus is, meaning that it is easily recognizable as belonging to its category. The lower the value of the time intensity parameter δ_j, the lower the time each stimulus needs to be categorized. The random structure of this model should be preferred when there is a low within–stimuli between–conditions variability, which can already indicate that the functioning of the stimuli does not change across conditions.

In Model 2, the within–stimuli between–conditions variability and the between–respondents variability are taken into account by specifying stimuli random slopes in the associative conditions and respondents random intercepts, respectively:

$$y_i = logit^{-1}(\alpha + \beta_i l_i + \alpha_{j[i]} + \beta_{k[i]} l_i + \epsilon_i), \tag{5.9}$$

with:

$$\beta_k \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_{\beta_{kl_1}} & \sigma^2_{\beta_{kl_1}, \beta_{kl_2}} \\ \sigma^2_{\beta_{kl_1}, \beta_{kl_2}} & \sigma^2_{\beta_{kl_2}} \end{pmatrix}\right), \tag{5.10}$$

$$\alpha_j \sim \mathcal{N}(0, \sigma^2_{\alpha_j}). \tag{5.11}$$

The random structure of Model 2 results in condition–specific stimuli parameters and overall respondents' parameters. The condition–specific easiness parameters b_{kl} are expressing the characteristics of the stimuli making them more easy to sort in the correct category according to the specific associative condition. The higher the value of b_k in one condition, the easier the stimulus in that condition. This means that the stimulus is sorted more easily when it is associated with the stimulus category of that condition than in the other one. Likewise, the condition–specific time intensity parameters δ_{kl} are expressing the characteristics of the stimuli for which they are obtaining faster responses according to the associative condition. If a stimulus shows a higher δ_k parameter in one condition than in the other, it means that it obtained faster responses when it was associated with the stimuli category in the former condition than in the latter one. For both the Rasch and the log-normal models, the condition–specific stimuli parameters allow for testing the associative condition in which the stimuli are presented is influencing their functioning. Moreover, by computing the difference between the condition–specific parameters it is possible to obtain a measure of the bias due to the associative conditions, hence providing information on the contribution of each stimulus to the IAT effect. The overall respondents' parameters, either ability θjl or speed τ_{jl}, are giving a general information regarding the respondents' performance. The random structure of

Model 2 should be preferred when a low within–respondents' between conditions variability is observed, which can already indicate that respondents' performance is not affected by the IAT associative condition.

In Model 3, used as a Null model, between–respondents variability and between–stimuli variability were specified as respondents and stimuli random intercepts, respectively:

$$y_i = logit^{-1}(\alpha + \beta_i l_i + \alpha_{j[i]} + \alpha_{k[i]} + \epsilon_i), \tag{5.12}$$

with

$$\alpha_j \sim \mathcal{N}(0, \sigma^2_{\alpha_j}), \tag{5.13}$$

$$\alpha_k \sim \mathcal{N}(0, \sigma^2_{\alpha_k}) \tag{5.14}$$

The random structure of Model 3 results in overall stimuli and respondents parameters. The random structure of this model should be preferred in those cases where there is a low variability at both the respondents and the stimuli levels. This lack of variability is already indicating that there is no IAT effect on respondents' performance as well as no variability in the stimuli functioning between conditions.

AN overview of the parameters that can be obtained from the random structures presented in this Section is illustrated in Table 5.1

Table 5.1: Accuracy and log-time models overview.

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

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

Chapter 6

Empirical application to IAT data

In the present chapter: (i) GLMMs have been applied to IAT accuracy responses to obtain estimates of Rasch model parameters; (ii) LMMs have been applied to IAT log-time responses to obtain estimates of log-normal model parameters; (iii) The relationship between the classic measure of the IAT effect (i.e., the *D-score*) and the models parameters obtained via the GLMMs and the LMMs has been investigated; (iv) The ability of model parameters and that of the *D-score* to predict a behavioral choice have been compared.

Two different IATs were considered, one for the assessment of attitudes towards African people and Caucasian people (Race IAT, Section 6.1), and one for the assessment of the preference for Dark Chocolate over Milk Chocolate (Chocolate IAT, 6.2).

Accuracy and log-time models were applied to both the Race IAT and the Chocolate IAT. Models were fitted with the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R (Version 3.5.1, R Core Team, 2018). The `implicitMeasures` (Epifania, Anselmi, & Robusto, 2019b) package was used for computing the IAT *D-score*.

6.1 Rasch gone mixed: A mixed model approach to the Implicit Association test

6.1.1 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%) self-identified as belonging to the Mediterranean ethnic group.

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

Data cleaning and *D-score*. Exclusion criteria based on both latency and accuracy responses were applied (Greenwald et al., 2003; Nosek et al., 2002). The algorithm *D1* in Greenwald et al. (2003) was used for computing the *D-score*. The difference was computed between the average response time in the BGWB and the WGBB condition: Positive scores stood for a possible preference for White people over Black people. In applying the LMMs, the latencies at the incorrect responses were used.

6.1.2 Race IAT: Results

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

Accuracy models. Concerning AIC, Log-Likelihood, and Deviance, Model A2 (AIC = 3784.43, Log-Likelihood = -1886.21, Deviance = 3722.43) performed better than Model A1 (AIC = 3786.51, Log-Likelihood = -1887.26, Deviance = 3774.51) and Model A3 (AIC = 3785.87, Log-Likelihood = -1888.93, Deviance = 3777.87). However, the latter one showed the lowest BIC value (3813.53, 3825.91, 3828.00, BIC values for Model A3, A2, and A1, respectively). Model A2 was chosen. This model provided overall participants ability parameters θ and condition–specific stimuli easiness parameters ($b_{\texttt{WGBB}}$ and $b_{\texttt{BGWB}}$). Results from Model A2 indicated a higher probability of correct response in the WGBB condition (*log-odds* $=$ 3.45, *SE* $= 0.12$) than in the BGWB condition (*log-odds* $=$ 2.07, *SE* $= 0.11$). Between–participants variability was 0.17. Between–stimuli variability in the WGBB condition ($\sigma^2 = 0.08$) was lower than the between–stimuli variability in the BGWB condition ($\sigma^2 = 0.15$). The correlation between stimuli variability in the two conditions was moderate ($r = .34$).

Stimuli easiness parameters are reported in Table 6.1. Overall, IAT stimuli tended to be easy stimuli. Stimuli tended to be easier in the WGBB condition than in the BGWB condition, where they showed a higher easiness variability. On average, object stimuli in the WGBB condition were the easiest stimuli, while negative words stimuli tended to be the least easy stimuli in the BGWB condition, immediately followed by positive words in the same condition. The difference in stimuli easiness parameters is reported in Table 6.1, as well. Object stimuli showed the lowest average easiness difference, while attribute stimuli, particularly positive word stimuli, showed the highest average difference between conditions. According to easiness difference, the stimuli giving the highest contribution to the IAT effect were *joy* (*Diff* = 1.85), *evil* (*Diff* = 1.82), *happiness* (*Diff* = 1.81), and *horrible* (*Diff* = 1.79).

Stimulus *bf3* showed the lowest difference between conditions (*Diff* $= 0.89$).

Table 6.1: Stimuli condition–specific easiness parameters (b_{jk}) and overall time intensity parameters (δ_j) - Race IAT

	b_{WGBB}	b_{BGWB}	$b_{\text{WGBB}} - b_{\text{WGBB}}$	δ_j		b_{WGBB}	b_{BGWB}	$b_{\text{WGBB}} - b_{\text{WGBB}}$	δ_j
Positive words					Negative words				
joy	3.53	1.69	1.85	0.02	evil	3.19	1.37	1.82	−0.01
happiness	3.48	1.67	1.81	0.01	horrible	3.56	1.77	1.79	0.05
pleasure	3.29	1.60	1.69	0.05	bad	3.11	1.58	1.53	0.03
peace	3.32	1.73	1.59	0.01	terrible	3.34	1.81	1.52	0.01
good	3.54	1.95	1.59	0.01	hate	3.34	1.85	1.50	0.01
laughter	3.54	2.03	1.52	0.09	failure	3.43	2.06	1.38	0.05
love	3.48	1.99	1.49	0.01	annoying	3.07	1.87	1.20	0.09
glory	3.42	1.99	1.43	0.08	pain	3.21	2.02	1.19	0.10
M (*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)
Caucasian American people					African American people				
wm3	3.61	2.04	1.57	−0.05	bm2	3.61	2.32	1.30	−0.08
wf3	3.66	2.29	1.36	−0.05	bf2	3.56	2.33	1.23	−0.06
wf2	3.59	2.46	1.12	−0.03	bf1	3.56	2.36	1.20	−0.04
wm2	3.48	2.44	1.04	0.03	bm1	3.52	2.42	1.10	−0.10
wf1	3.59	2.57	1.02	−0.05	bm3	3.58	2.51	1.07	−0.09
wm1	3.28	2.28	1.01	−0.02	bf3	3.36	2.47	0.89	−0.05
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”: Caucasian American female face, “wm”: Caucasian American male face, “bf”: African American female face, “bm”: African American male face; WGBB: White-Good/Black-Bad condition; BGWB: Black-Good/White-Bad condition. Rows are ordered by decreasing values of $b_{\text{WGBB}} - b_{\text{WGBB}}$

Log-normal models. The three log-time models were compared between each other. Model T2 produced aberrant estimates (i.e., correlation between the stimuli random slopes equal to 1). Model T1 (AIC = 4399.66, BIC = 4448.06, Log-Likelihood = -2192.83, Deviance = 4385.66) performed better than Model T3 (AIC = 4762.63, BIC = 4797.20, Log-Likelihood = -2376.32, Deviance = 4752.63). Model T1 was chosen. This model provided condition–specific participants' speed parameters (τ_{WGBB} and τ_{BGWB}) and overall stimuli time intensity parameters δ. Responses in the WGBB condition tended to be faster ($\beta = -0.43$, $SE = 0.02$) than responses in the BGWB condition ($\beta = -0.15$, $SE = 0.03$). The between-stimuli variability was particularly low ($\sigma = 0.003$), while the between-participants variability was slightly higher in the BGWB condition ($\sigma = 0.05$) than in the WGBB one ($\sigma = 0.02$). The correlation between respondents' variability in the two conditions was strong ($r = .63$).

Stimuli time intensity parameters δ are reported in Table 6.1. Attribute stimuli required more time to get a response, while object stimuli were the ones requiring less time, with Black people faces inducing the fastest responses. Black male faces required less time to obtain a response than Black female faces did, while this pattern was not observed for White people faces. Three of the positive attribute stimuli (*pleasure*, *glory*, *laughter*) showed time intensity estimates higher than the estimates of the stimuli belonging to the same category. Also three negative words (*failure*, *annoying*, *pain*) showed a higher time intensity estimates than the other negative words. Object stimuli tended to have similar time intensity estimates.

Regression model: *D-score*. A *speed-differential* measure was computed as the difference between speed parameters in the BGWB condition and the WGBB condition. Negative values indicated a respondent faster in the BGWB condition than in the WGBB condition. Pearson's correlations were computed between participants' ability, condition–specific speed parameters and *speed-differential*. Participants' ability poorly and positively correlated with speed in the BGWB condition ($r = .13$, $p = .316$), while it correlated negatively and poorly with the *speed-differential* ($r = -.14$, $p = .279$). Ability moderately correlated with the speed parameters in the WGBB condition ($r = .32$, $p = .011$). Participants' ability and *speed-differential* were regressed on the *D-score*, and backward deletion was used to investigate the linear combination of predictors accounting for the higher proportion of explained variance. Backward deletion kept both the predictors in the model, which accounted for about 70% of the total variance (*Adjusted* $R^2 = .78$, $F(2, 59) = 106.3$, $p < .001$). *Speed-differential* strongly and positively predicted *D-score* ($\beta = 1.93$, $t(59) = 13.88$, $p < .001$): Ability negatively predicted the *D-score* ($\beta = -0.18$, $t(59) = -2.48$, $p = .016$).

6.2 Filling the attitude-behavior gap: A Rasch modeling of the Implicit Association Test

Study 2 aimed at replicating the models from Study 1 on a different IAT, and at investigating the relation between model parameters and behavioral outcomes. An IAT for the assessment of the implicit preference towards Dark or Milk chocolate was hence developed. The use of a Chocolate IAT allowed for the inclusion of both a behavioral choice and of the explicit investigation of the chocolate preference.

6.2.1 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 by ? (?). Specifically, twenty-six attribute stimuli (13 positive words and 13 negative words) and seven chocolate images were used. The images of chocolate were modified to represent either Milk or Dark chocolate, resulting in 14 object stimuli.

Respondents were presented with 60 trials in the Dark-Good/Milk-Bad (DGMB) condition, as well as 60 trials in the Milk-Good/Dark-Bad (MGDB) condition. Participants did not receive any feedback on their performance, and they were instructed to be as accurate and fast as they could. Participants' explicit chocolate preferences were investigated with two items (i.e., *"How much do you like Milk/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 piece of Dark or Milk chocolate as a reward for their participation. The experimenter registered participants' choices after they left the laboratory.

Data cleaning and *D-score*. The same data cleaning procedures employed in Study 1 were applied. Algorithm *D4* in Greenwald et al. (2003) was used for the *D-score* computation. The difference was computed between the average response time in the MGDB condition and in the DGMB condition: Positive scores stood for a possible preference for Dark chocolate. In applying the LMMs, the latencies at the incorrect responses were used.

6.2.2 Chocolate IAT: Results

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

Accuracy models. Model A2 from Study 1 was applied to Chocolate IAT accuracy responses and compared against the other Accuracy Models. Model A1 failed to converge, while Model A2 (AIC = 3625.58, Log-Likelihood = -1806.79, Deviance = 3613.58) performed better than Model 3A (AIC = 3627.71, Log-Likelihood = -1809.85, Deviance = 3619.71). Nonetheless, Model A3 showed a lower value of BIC than Model A2 (3656.07, 3668.13 for Model A3 and Model A2, respectively). Model A2 was chosen. The model resulted in the estimation of overall respondents' ability θ 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 DGMB (*log-odds* $= 2.61$, *SE* $= 0.10$). Participants showed a high variability ($\sigma^2 = 0.33$). Between-stimuli variability was higher in the MGDB condition ($\sigma^2 = 0.21$) than in the DGMB condition ($\sigma^2 = 0.01$). Stimuli variability in the two conditions were weakly correlated ($r = .20$).

Stimuli easiness parameters are reported in Table 6.2. Stimuli tended to be easier in the MGDB condition than in the DGMB condition. On average, positive words and images of Milk chocolate had a similar easiness estimate in the MGDB condition. Milk chocolate stimuli showed the highest easiness difference between the two conditions, while the lowest difference was found for negative words stimuli. Positive attribute stimuli, like *joy* (*Diff* = -1.40) and *happiness* (*Diff* = -1.39), contributed the most to the IAT effect, while negative attribute stimuli tended to give a lesser contribution, despite some exceptions, like *hate* (*Diff* = -1.26) and *failure* (*Diff* = -1.25). *Dark 5* was the object stimulus contributing the most to the IAT effect (*Diff* = -1.38), while *agony* was the stimulus that gave the least contribution (*Diff* = 0.08).

Table 6.2: Stimuli condition–specific easiness parameters (b_{jk}) and overall time intensity parameters (δ_j) - Chocolate IAT

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

Note: DGMB: Dark-Good/Milk-Bad condition; MGDB: Milk-Good/Dark-Bad condition; Difference: Difference between DGMB and MGDB condition. Rows are ordered by absolute decreasing values of $b_{\text{DGMB}} - b_{\text{DGMB}}$

Log-normal models. Model T1 from Study 1 was applied to log-time responses of the Chocolate IAT and compared against Model T2 and Model T3. Model T2 produced aberrant estimates. Model T2 (AIC = 7159.23, BIC = 7208.87, Log-Likelihood = -3572.62, Deviance = 7145.23) performed better than model T3 (AIC = 7856.45, BIC = 7891.91, Log-Likelihood = -3923.23, Deviance = 7846.45). Model T1 was chosen. The model resulted in overall stimuli time intensity parameters δ and respondents' condition–specific speed parameters ($\tau_{\texttt{MGDB}}$ and $\tau_{\texttt{MGDB}}$). Responses tended to be faster in the MGDB condition ($\beta = -0.36$, *SE* $= 0.02$) than in the DGMB condition ($\beta = -0.12$, *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 ($r = .40$).

Stimuli time intensity parameters δ_k are reported in Table 6.2. Object stimuli tended to require less time to get a response than attribute stimuli did. Specifically, images of Dark chocolate were the stimuli requiring, on average, the least amount of time for getting a response. Negative words stimuli were the stimuli requiring more time to get a response, with stimulus *evil* requiring the longest time ($\delta = 0.10$).

Regression model: *D-score*. A *speed-differential* was computed as the difference between participants' speed parameters in the MGDB condition and the DGMB condition. Positive values indicated faster respondents in the DGMB condition. Pearson's correlation was computed between participants' ability, condition–specific speed parameters, and *speed-differential*. Participants' ability correlated neither with participants' speed in the DGMB condition ($r = .07$, $p = .559$), nor with participants' speed in the MGDB condition ($r = .18$, $p = .530$), and nor with *speed-differential* ($r = .07$, $p = .529$).

Speed-differential, participants' ability, and explicit preferences for Milk and Dark chocolate were regressed on the *D-score*. After backward deletion, the model including only the *speed-differential* accounted for the higher proportion of explained variance (*Adjusted* $R^2 = .90, F(1, 72) = 685.60, p < .001$). *Speed-differential* strongly and positively predicted the *D-score* ($\beta = 2.46$, $t(72) = 26.18$, $p < .001$).

Regression model: Behavioral outcome. The ability of the models parameters and that of the *D-score* to predict a behavioral choice was investigated and compared. Dark chocolate choice (DCC) was labeled as 0 and Milk chocolate choice (MCC) was labeled as 1. Explicit evaluations of Milk and Dark chocolate were not included since they were asked right before the behavioral choice. Collinearity was found between *D-score*, *speed-differential*, and condition–specific speed parameters (Variance Inflation Factors, *VIF*s > 10). No collinearity was found between other predictors (i.e., condition–specific speed, participants' ability, food habits, *VIF*s < 4.00). Three models were specified, including only one of the relevant predictors at a time (*D-score*, *speed-differential*, the two condition–specific speed parameters). After 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) were used as criteria to decide which of the predictors best accounted for the actual choice. After backward deletion, only *D-score* (*Deviance* $= 84.70$, *df* $= 72$, *Nagelkerke* $R^2 = 0.19$), *speed-differential* (*Deviance* $= 84.45$, *df* $= 72$, *Nagelkerke* $R^2 = 0.20$) and condition–specific speed parameters (*Deviance* $= 83.88$, *df* $= 71$, *Nagelkerke* $R^2 = 0.20$) remained in their respective models. *D-score* (*log-odds* $= -2.03$, $z(72) = -3.36$, $p < .001$) and *speed-differential* (*log-odds* $= -5.02$, $z(72) = -3.51$, $p < .001$) negatively predicted the chocolate choice. Speed in MGDB (*log-odds* $= -5.99$, $z(71) = -3.02$, $p = .003$) negatively predicted the choice, while speed in DGMB (*log-odds* $= 4.73$, $z(71) = 3.19$, $p < .001$) positively predicted it. The model including the *D-score* showed a general accuracy of 66%, a DCC accuracy of 70%, and an MCC accuracy of 61%. The model including the *speed-differential* correctly identified 68% of chocolate choices, 72% of DCCs, and 61% of MCCs. The model with the condition–specific speed parameters showed a general accuracy of 72%, a DCC accuracy of 74% of cases, and a MCC accuracy of 68%.

6.3 Discussion and Conclusions

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

The fine-grained analysis at the stimuli level allowed for a deeper understanding of the meaning of the IAT measure, for example by giving the chance of investigating the stimuli that either were not representative of their category or were not contributing to the IAT effect. If a stimulus was hardly recognizable as belonging to its alleged category, it would have resulted in a lower easiness parameter, as well as a higher time intensity parameter. Additionally, this stimulus was expected to have similar condition–specific easiness parameter estimates, which suggested a low contribution to the IAT effect. The information available at the stimuli level can also be used for highlighting the stimuli giving the highest contribution to the IAT effect. These stimuli can be selected and employed for the design of briefer, but still highly informative IATs.

These models also resulted in detailed information on respondents' accuracy and speed performance. Indeed, understating how respondents are behaving during the IAT administration is crucial for getting a deeper comprehension of the measure obtained and of the factors that might influence it. In both the considered IATs, respondents' accuracy performance was not affected by the IAT conditions, while their speed performance was. Drawing on this result, it can be speculated that the IAT effect is mostly due to respondents' slowdown in one condition while the accuracy performance remains consistent between the two conditions.

The *D-score* was strongly related with respondents' *speed-differential*, while it was related with their ability only when the built-in correction was included. The administration procedure of the IAT itself (i.e., providing or not feedback for incorrect responses) might play a role in influencing the IAT effect, at least regarding the *D-score*. Nonetheless, this result may be due to different concurring factors (e.g., stimuli employed, number of stimuli), and it might deserve a more thorough and systematic investigation.

While the *speed-differential* did not outperform the *D-score* in predicting the behavioral choice, the condition–specific speed parameters did. Different explanations should be considered for this result. Firstly, a differential measure, like the *D-score* or the *speed-differential*, might conceal important components of the implicit preference that can be acknowledged when the condition–specific parameters are considered. Nonetheless, when the single components of the *D-score* (i.e., average response time in each associative condition) were used to predict the behavioral choice, the accuracy of the prediction was still outperformed by the condition–specific speed parameters performance. Since the *D-score* is obtained by averaging across all the trials in the two associative conditions, it can result in biased estimates of the implicit preference (Wolsiefer et al., 2017). Thus, the *D-score* computation strategy itself could have affected its prediction ability. Finally, the *D-score* is computed including the inflated time for the incorrect responses, while condition–specific speed parameters are estimated using the actual response time for the error responses. These elements taken together might bring to a more accurate inference of the implicit preference from the condition–specific speed, resulting in a better prediction of the choice.

So far, information gained from accuracy responses (participants ability and stimuli easiness) and from time responses (participants' speed and stimuli time intensity) are not combined together. Combining these information, for example with the Hierarchical Model for Accuracy and Speed (van der Linden, 2007), might provide further insights on the IAT general functioning. At the stimuli level, it would be possible to investigate whether particularly easy stimuli are also the fastest stimuli, or, vice-versa, whether more difficult stimuli are also the stimuli requiring more time to get a response. If not, it would probably mean that the responses to a particularly difficult stimulus are mostly given by chance, hence resulting in higher proportion of incorrect responses and faster responses.

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

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

Chapter 7

Making it comprehensive: Mixing the Implicit

To the best of our knowledge, no attempts of modeling the SC-IAT responses with a Rasch modeling approach have been made so far. Moreover, it appears that no attempts to concurrently model IAT and SC-IAT responses under one comprehensive model have been made so far.

In the present study, we aim to fill these gaps by presenting a modeling framework able to simultaneously account for the sources of dependency between and within IAT and SC-IAT responses, and to provide Rasch model and log-normal model estimates. To pursue these aims, LMMs with appropriate random structures have been specified for modeling the IAT and SC-IAT responses. Three levels of modeling complexity have been considered. At a first level, the IAT and SC-IAT observations have been modeled separately with distinct models (Single measure models). At a second level, both the IAT and the SC-IAT observations have been modeled concurrently, but only the between–measures variability has been considered (Type measure models). Finally, at the third and most complex level, the IAT and the SC-IAT observations have been modeled together by accounting for the associative conditions of each implicit measure (Implicit condition models). The relationship between the parameters obtained through these models and classic IAT and SC-IAT *D-score*s has been investigated as well. Finally, the ability of the Rasch and log-normal parameter estimates and that of the classic *D-score*s to predict a behavioral outcome have been investigated and compared. It is expected that the Implicit condition model would result in more reliable estimates, as expressed by their relationship with the classic scores and the prediction of the behavioral outcome.

Chapter 8

The feedback effect on implicit measures performance

Built in vs no built-in

Chapter 9

General discussion

Last note

I just hope the end will come soon, silently, and painless.

References

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

Anselmi, P., Vianello, M., & Robusto, E. (2013). Preferring thin people does not imply derogating fat people. A Rasch analysis of the implicit weight attitude. *Obesity*, *21*(2), 261–265. doi: 10.1002/oby.20085

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)

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688.

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.

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

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

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

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

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)

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

Conrey, F., Gawronski, B., Sherman, J., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. 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. , *39*(12), 1–28.

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

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. (2019a, 10 30). Dscoreapp: An user-friendly web application for computing the implicit association test d-score. *Journal of Open Source Software*, *4*(42), 1764. Retrieved from `http://fisppa.psy.unipd.it/DscoreApp/` doi: 10.21105/joss.01764

Epifania, O. M., Anselmi, P., & Robusto, E. (2019b). implicitMeasures: Computes the Scores for Different Implicit Measures [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=implicitMeasures` (R package version 0.1.0)

Epifania, O. M., Robusto, E., & Anselmi, P. (2020, 2). Implicit social cognition through years: The Implicit Association Test at age 21. Retrieved from `$https://advance.sagepub.com/articles/Implicit_social_cognition_through_years_The_Implicit_Association_Test_at_age_21/11914416$` doi: 10.31124/advance.11914416.v1

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.

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., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of personality and social psychology*, *79*(6), 1022. doi: 10.1037//0022-3514.79.6.1022

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

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. (1960). *Many-Facet Rasch measurement.* Chicago, IL: MESA Press.

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., & 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

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

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

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.

Software, C. (2011). *Inquisit 3.0.6.0.* `https://www.millisecond.com`.

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)

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

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

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

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.

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

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.