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

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

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

Advisor: Prof. Egidio Robusto

Co-Advisor: Prof. Gianmarco Altoè

Ph.D. Candidate: Ottavia M. Epifania

Academic Year: 2019/2020

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

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

This page is NOT intentionally left blank. Please help.

Contents

Preface

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The structure of the thesis is outlined.

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

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

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

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

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

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

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

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

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

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

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

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

The log-normal model is introduced in the first section of Chapter **??** 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.

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

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

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

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

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

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

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

The typical scoring methods of both the IAT and the SC-IAT have been presented, and their predictive ability in respect to a behavioral outcome has been investigated and compared with that provided by new scoring methods (Chapter **??**). 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.

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

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

Chapter **??** 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. The IAT and the SC-IAT data have been modeled separately with the (G)LMMS of Chapter **??** for obtaining the Rasch model and the log-normal estimates from each of them singularly. This was done for mainly two reasons. Firstly, to investigate the soundness of the proposed approach for modeling measures other than the IAT. Secondly, to investigate whether and how model estimates change if the within–respondents between–measures variability and the within–stimuli between–measures variability are not accounted for.

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

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

Finally, Chapter 1 summarizes the findings of all other chapters, and draws general conclusions based on the evidence reported in all the studies.

Chapter 1

Conclusions

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

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

The sound path

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

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

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

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

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

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

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

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

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

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

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

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

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

Anselmi, Vianello, Voci, & Robusto, 2013).

This result is further corroborated by the low difference between the condition–specific estimates of the category *Black*. As such, the easiness of categorization of these stimuli did not change much depending on the evaluative dimension with which they shared the response key. It can be speculated that Black people were neither strongly associated with negative attributes nor with positive ones, and that the resulting IAT effect was mostly driven by the evaluations made on White people faces.

The best fitting model for the log-time responses was the one allowing for the multidimensionality at the level of the respondents, while stimuli were centered at 0. This model resulted in the estimation of condition–specific respondents' speed parameters and overall stimuli time intensity estimates. The best fitting model indicated that there was a high within–respondents between–conditions variability along with a low within–stimuli between–conditions variability. This implies that respondents' performance changed between the two associative conditions, while the functioning of the stimuli remained the same between conditions. In other words, the time each stimulus required for getting a response did not change according to the stimuli category with which they shared the response key. The overall time intensity estimates can inform about the within–stimuli categories variability, and hence about stimuli hetereogeneity. Specifically, stimuli displaying a time intensity estimate too far away from the time intensity estimates of the other stimuli belonging to the same category should be replaced to reduce both the within–categories variability and the between–stimuli variability.

The condition–specific respondents' speed estimates allowed for delving deeper on the association(s) driving the IAT effect. Additionally, they provided a differential measure similar to the D score that expresses the bias on the speed performance due to the effect of the associative conditions. This differential measure can be used for further analysis, such as the prediction of behavioral outcomes.

The first study brought evidence in favor of the usefulness and feasibility of the proposed modeling framework for the analysis of the IAT data. However, neither the usefulness of the information at the stimuli level nor that at the respondents' level were tested. If the stimuli estimates provided by these models inform on the stimuli giving the highest contribution to

the IAT effect, it should be possible to isolate and select them for obtaining better performing IATs. Indeed, selecting only the stimuli giving the highest contribution to the IAT effect would reduce the stimuli heterogeneity and consequently the across-trials variability. If this is true, even the *D* score would result in a more reliable measure of the implicit construct under investigation. Moreover, the Rasch and log-normal model estimates obtained from the application of the (G)LMMs to IAT responses are less affected by sources of error variance than the *D* score is. Consequently, they result in a better inference of the construct under investigation and, as such, they potentially lead to a better prediction of a behavioral outcome.

In the second study, the two above-mentioned points were directly tested by using an IAT for the assessment of the implicit preference for Dark or Milk chocolate (Chocolate IAT). Both the accuracy and the log-time models were replicated on this data set. Condition–specific stimuli easiness estimates and overall ability estimates of the Rasch model, and condition–specific respondents' speed estimates and overall stimuli time estimates of the log-normal model were obtained. Also in this case, all stimuli tended to be easier in one condition over the other. Specifically, stimuli tended to be easier in the Milk-Good/Dark-Bad condition, and *Good* evaluative attributes were the stimuli having the greatest impact on the IAT effect. Given this pattern of results, it is possible to speculate that it was more the liking for Milk chocolate than the dislike for Dark chocolate that drove the performance at the IAT.

By using dark and milk chocolate as target objects of the IAT, it was possible to reward the participation of the respondents with a free bar of dark or milk chocolate. Obviously, the free bar of chocolate was not just a reward for the respondents but also the behavioral task of the experiment. The choice was registered by the experimenter, and it was used for investigating the predictive ability of the model estimates and that of the *D* score. Specifically, the choice was predicted by both the differential measures and the linear combination of the their single components in different logistic regressions. Backward deletion and the accuracy of the choice prediction provided by the models were used to determine the predictors best accounting for the observed chocolate choice. The log-normal speed estimates outperformed the *D* score in the prediction of the behavioral outcome. The lower predictive ability of the *D* score was observed both in comparison to its own linear components (although they explained

a lower proportion of variance) and in respect to both the linear combination of the condition–specific speed estimates and their *speed-differential*. The *D* score and its linear components do include uncontrolled sources of error variance due to the multiple sources of random variability in the IAT data. On the other hand, these sources of variability are accounted for in the speed estimates. Since error variance is most unlikely related to behaviors (Meissner, Grigutsch, Koranyi, Müller, & Rothermund, 2019), it should not be surprising to find a lower predictive ability of the *D* score. The *speed-differential* resulted in a slightly lower predictive ability than that provided by the linear combination of its single components.

The second study also investigated the ability of the condition–specific easiness estimates to pinpoint the stimuli giving the highest (lowest) contribution to the IAT effect. Since across-trial variability due to the stimuli heterogeneity is one of the factors that mostly affects the computation of the IAT *D* score, the reduction of the stimuli heterogeneity by selecting a specific pool of stimuli should provide more reliable *D* scores. By selecting only the stimuli providing the highest contribution to the IAT effect or the ones providing the least contribution to the IAT effect, the number of trials was reduced to $1/3$ of the original starting trials pool. Two additional *D* scores were computed, one on the data set including the stimuli giving the highest contribution to the IAT effect, and one on the data set including the least informative stimuli. The *D* score computed on the high informative stimuli data set did show a slightly better performance than the *D* score computed on the low informative stimuli. This result brings further evidence on the sensitivity of the *D* score to the across-trial variability due to the heterogeneity of the stimuli, at the point that it does not even matter whether the highest informative stimuli or the lowest informative ones are selected, as long as the variability is reduced.

The (G)LMMs approach showed its feasibility and appropriateness also for modeling SC-IAT data within a Rasch framework. However, this result has to be contextualized into the specific context of this thesis, where both the IAT and the SC-IATs were administered together. By separately analyzing the data from the three implicit measures, there are still sources of error variance that are left free to bias the estimates of the parameters. However, if the SC-IAT is administered as a stand-alone measure, a Rasch parametrization of its accuracy

and log-time responses can be easily obtained with the modeling framework presented in this work.

The comprehensive modeling approach of the IAT and the SC-IAT highlighted an IAT effect on both accuracy and speed performance of the respondents in all implicit measures. This result might indicate that, despite respondents slowed down in one condition, their accuracy performance is still impaired in that condition. Consequently, it is not surprising to find ability estimates of both the Single measure models and the Comprehensive models in predicting the typical score of each implicit measure. Indeed, typical scoring of the IAT and the SC-IAT are based on both time and accuracy responses (i.e., error responses are replaced with the average response time added with a penalty). The higher the number of incorrect responses, the higher the number of trials whose response time is replaced with the inflated one, and, consequently, the higher the average response time. It logically follows that, if in one condition both the speed performance and the accuracy performance are impaired, the average response time will be higher due to the combined effect of the slower response times and the inflated error response times, resulting in a higher difference between the associative conditions and in a larger effect size. Nonetheless, the ability estimates had a smaller effect size in the prediction of the typical scoring than the speed estimates.

The difference in respondents' performance between the associative conditions due to their slowing down and/or to a higher number of mistakes might be ascribable to just a small set of stimuli. In this case, the difference is not entirely related to automatic evaluative associations but also to the peculiarities of the task. Therefore, understanding how and why the estimates of some stimuli are far away from the those of the stimuli belonging to the same category becomes of particular relevance for getting a better and deeper understanding of the measure obtained, and of the inferences that can be reasonably done. If a stimulus is correctly responded but requires a longer time, it influences the average response time, hence skewing the result. If a stimulus is incorrectly responded, its response time is replaced by the average response time in that condition added with a penalty. Either way, the effect size of the *D* score will be artificially inflated by the response time of just some of the stimuli, and the inferences based on that should be taken with caution. Nonetheless, in considering the results on the

stimuli functioning, it was not possible to rule out the effect of the associative conditions.

The typical scores of both SC-ATs were always cut out in the prediction of the behavioral outcome. This result held for both the typical differential scores and the linear combination of their single components. If one was called to draw conclusions on the contribution of the SC-IATs to the prediction of behavioral outcomes, he/she would have probably inferred that the SC-IATs do not give any contribution to the choice prediction. Consequently, only the measure obtained from the IAT would have been considered as relevant for predicting behaviors.

Indeed, also the differential measures obtained from the model parameters estimates, both with the Single measures models and the Comprehensive models, pointed in the same direction as the typical scoring methods. Only the differential measures obtained from the IAT condition–specific speed estimates have been found to predict the choice, while the differential measures obtained from the estimates of the SC-IAT did not contribute in predicting the choice. However, when the linear combination of the condition–specific speed estimates of each implicit measure was used for predicting the choice, the speed in the Dark-Good condition of the Dark SC-IAT entered and remained in the model.

These results point at the risks of using the *D* score as a measure of the implicit construct under investigation. As already discussed, the *D* scores, as well as their linear components, are affected by sources of uncontrolled error variance resulting from the data structure itself. The administration of multiple measures to the same respondents generates further sources of variability and dependencies. Consequently, these scores result in biased estimates which, in this specific case, are not able to highlight the contribution of each implicit measure in the prediction of a behavioral outcome. The conclusions drawn from such scores should hence be taken with cautions.

Moreover, the results obtained on the choice prediction from both the study in Chapter **??** and that in Chapter **??** highlighted the issue related to the use of differential measures. In both cases, regardless of the implicit measure under consideration or the modeling framework used, differential measure are less accurate in predicting the behavioral outcome than the linear combination of their respective single components. This result is more evident for

the speed estimates of the log-normal model. The differences between the typical scores of implicit measures and their linear components is less evident, probably because the prediction is already affected by other sources of error variance. In Chapter **??**, the model including the single components of the *speed-differential* was the one resulting in the highest accuracy of prediction of the Milk chocolate choice. Milk Chocolate Choice was disregarded by the *D* score, its linear components, and the *speed-differential*. In Chapter **??**, the model including the linear combination of the condition–specific speed estimates of each implicit measure was the only one able to highlight the contribution of the speed of the Dark-Good condition of the Dark SC-IAT. This model did not result in a higher predictive accuracy than the others, but it did explain a higher proportion of variance of the choice. Besides, it made possible to gain a better understanding of the processes underlying the choice.

In the former case, differential measures did not provide a good prediction of one of the possible outcomes. In the latter one, differential measures were not able to identify the contribution of the SC-IAT in predicting the choice. The speed in only one of the conditions of the SC-IAT was found to contribute to choice prediction. The differential measure computed between the speed of the two conditions of the Dark SC-IAT might have confounded their importance and relevance for choice prediction, pointing at a null contribution of the Dark SC-IAT. Remarkably, the contribution of the Dark SC-IAT was completely lost when the linear components of the typical scoring were used.

The lack of predictive ability of differential measures might be due to the nature of differential measures themselves, which is confounding the contribution of each single component used for the computation of the single score. The computation of differential measures results in reliable scores only when the two quantities used for the computation have the same weight in determining the final score. This can be true only if a series of assumption are met, and, in the IAT case, this rarely happens (Fiedler, Messner, & Bluemke, 2006). Firstly, the two target categories are assumed to give the same exact contribution to the IAT effect. In other words, the liking for one of the target categories has to be as strong as the dislike for the opposite category. This logically implies that the zero point stands for the absence of any positive or negative attitudes toward both target objects. Secondly, also the evaluative

dimensions and the target objects are assumed to have the same impact on the IAT effect. This assumption is in line with the idea of treating the stimuli as a fixed factor, which implies that they all have the same impact on the observed scores. However, as extensively discussed in the first chapter, considering stimuli as fixed factors in the IAT case is a stretch, and the distinct contribution of each stimulus to the IAT effect should be taken into account. Finally, systematic and unsystematic sources of variability are assumed to affect respondents' performance across the two conditions in the same way.

The information on respondents' performance and stimuli functioning provided by the modeling framework proposed in this thesis can be used for verifying the assumptions that have to be met for the computation of reliable and meaningful differential measures. Results on the stimuli functioning clearly suggest that each stimulus does give a different contribution to the IAT effect, and that their variability differently affect the final score. For instance, all studies highlighted a higher time intensity estimate for the attribute stimuli than for the image stimuli. The former ones required less time for getting a response than the latter ones. Additionally, in some cases image stimuli tended to be easier than attribute stimuli. These results clearly point at a different processing of the stimuli according to their type (attributes or images), and, blatantly, they cannot have the same effect on the observed responses.

Moreover, the contribution of the stimuli to the IAT effect and the relationship between the respondents' condition–specific speed estimates and the *D* score suggest that the differential score is mostly driven by the performance in one of the two associative conditions. Consequently, it seems bold to assume that the liking for one of the target categories is as strong as the dislike for the other one. Especially in the IAT case, which rests its measure on the juxtaposition between two objects, there might be cases in which the preference (dislike) for one object is extremely strong, while the contrasting object is not related to any particular positive or negative evaluation. Therefore, it can be assumed neither that attitudes towards the two contrasting objects have the same importance for the final score, nor that stimuli are processed in the same way and have the same impact on the final differential measure.

Regarding the violation of the last assumption, the one regarding the sources of systematic variability affecting the two conditions in the same way, the description of the fully-crossed

structure of implicit measures provided in the first chapter should have already clarified why this assumption is not meant to hold. Moreover, also the conceptualization of the ReAL model presented in Chapter **??**, according to which different controlled processes can differently affect respondents' performance in the two associative conditions, makes hard to believe that this assumption could hold. Finally, the results on respondents' ability and speed performance obtained from the Rasch and log-normal modeling of the implicit measures do point to a difference in respondents' variability between the conditions. As a consequence of the violation of these assumptions, differential measures might not represent the best choice for expressing the implicit psychological construct assessed by implicit measures.

Given that the results on the relationship between model estimates and typical scoring methods and those on the choice prediction are almost identical for the estimates obtained with the Single measure models and those obtained with the Comprehensive model, one might be wondering about the advantages of using the latter approach over the former one. The former approach does result in measure–specific stimuli estimates, which inform about the functioning of the stimuli in each implicit measure. Conversely, the Comprehensive model results in overall stimuli estimates across implicit measures, hence providing a general information of stimuli functioning across measures. However, the apparent advantage of the Single measure model of providing measure–specific stimuli estimates is also its major shortcoming, as already discussed. By not addressing the between–measures variability, the new sources of error variance related to the administration of multiple implicit measures are left free to bias the estimates. Moreover, since the estimates are obtained from separate and independent models, they cannot be compared between each other. A comparison between the respondents' performance in the implicit measures is meaningless if not dangerous in terms of inferences that can be made.

The fair path

The fair path appears to be in clear antithesis with what has been said so far about typical scoring of implicit measures. However, effect size measures are still the most common ways for

scoring implicit measures data, both when administered as stand-alone measures and when they are administered together. The resulting scores are then used for further analyses and/or for comparing the performance of the different implicit measures in respect to some criteria (e.g., prediction of behavioral outcomes). However, the differences in both the administration and the scoring procedures of implicit measures such as the IAT and the SC-IAT might directly affect the score obtained at each of them. If these scores are then used for comparing the IAT and SC-IAT performance on different criteria, the comparison might result affected by artifacts which are not directly related to the goodness of the implicit measures itself but to elements of minor importance. How one can be sure that the lesser predictive ability in respect to a behavioral outcome provided by the SC-IAT is truly ascribable to the measure itself and not to some minor features? By providing easy-to-compute and easy-to-interpret effect size measures with which typical users of these implicit measures are more familiar with, the approach presented in the fair path might help in answering this question or, at the very least, in fostering a fairer comparison between the IAT and the SC-IAT. Summarizing, the fair path was aimed at providing rigorous and comparable scoring methods for different implicit measures without moving apart from the typical approach.

While it is true that different implicit measures do have features that make them unique, there are features that can be aligned in both their administration and their scoring. The alignment of these differences allows for a fairer comparison between the performance of implicit measures. This leads to mainly two advantages. Firstly, the performance of the respondents on different measures can be reasonably compared, and secondly the results on the comparison between implicit measures performance in respect to different criteria can be mostly ascribed to the measure and not to other artifacts.

The new scoring methods that have been implemented do not necessarily result in a higher accuracy of the prediction. They do point to a higher predictive ability of the IAT in respect to that of the SC-IAT. In this case, the better performance of the IAT can be more easily pinned to the measure itself and not to artifacts due to the differences of scoring and administration procedures. Moreover, by taking out the role of the scoring in potentially influencing the results, it is possible to make more accurate speculations on the reasons why the IAT does

show a better performance than the SC-IAT. In the study reported in Chapter **??**, the higher predictive of the IAT in respect to the SC-IAT might be due to the dichotomous nature of the choice, which is more in line with the comparative measure provided by the IAT than the absolute one provided by the SC-IAT.

Limitations and future directions

The modeling framework introduced in this thesis provides interesting and useful information on the functioning of different implicit measures, concerning both the respondents and the stimuli. For instance, it was possible to pinpoint the stimuli that gave the highest contribution to the IAT effect. This information can be further used for getting a better understanding of the automatic association(s) implicated in the performance at the IAT. Moreover, the information at the stimuli level help in reducing the across-trial variability by selecting only the most informative stimuli. By doing so, the administration time of the IAT can be reduced. At the respondents' level, it was possible to shed a new light on the components included into the *D* score, and to obtain a better inference on the implicit constructs under investigation.

However, the information yielded from the accuracy responses completely ignores the information yielded from the log-time responses, and vice versa. As such, important relationship between the responses might be lost. For instance, it is not possible to know whether an extremely easy stimulus (i.e., a stimulus that obtains a high proportion of correct responses) is as such because respondents tend to spend a high amount of time on it before giving a response or whether it also obtains fast responses. In the latter case, the stimulus can be considered as a good functioning one from both an accuracy and a time perspective. Similarly, if a stimulus has a low time intensity estimate (i.e., it obtains fast responses) combined with a low easiness estimate (i.e., it obtains a high proportion of incorrect responses), it should not be considered as a good functioning stimulus.

The separate modeling of accuracy and time responses assumes that the distributions of these variables are determined by different parameters, which are in turn generated by different processes (van der Linden, 2006). The accuracy and speed performance of one respondent

is constrained by a speed-accuracy trade-off. Once the speed-accuracy trade-off is set, the response time distribution of the respondent is solely determined by his/her speed. Similarly, the distribution of the accuracy responses only depends on the respondent's ability. However, when a population of respondents is considered, it is not possible to assume a single speed-accuracy trade-off, and a dependency between the accuracy and time responses should be expected (van der Linden, 2006, 2007). The relationship between the parameters governing the accuracy and speed performance can be understood at a second level of modeling, as illustrated in the hierarchical model by van der Linden (2007). As the name suggests, the hierarchical model posits two levels of modeling. At a first level, the accuracy and log-time responses are modeled separately. An IRT model is used for modeling accuracy responses, while the log-normal model is used for modeling the log-time responses. Each model yields stimuli and respondents' parameters explaining the accuracy and log-time responses. At a second level, two models are assumed to explain the relations between the respondents' parameters (i.e., *population model*) and the stimuli parameters (i.e., *item-domain model*). The population model assumes a multivariate normal distribution to describe the population from which the respondents are drawn. The multivariate distribution is defined by the respondents' parameters obtained from the accuracy and log-time models. The item-domain model describes the domain (population) of the items from which the items are drawn by assuming a multivariate normal distribution defined by the stimuli parameters obtained from the accuracy and log-time models.

Undoubtedly, the second level of modeling introduced by van der Linden (2007) would provide further insights on the functioning of implicit measures, concerning both the stimuli and the respondents. Nonetheless, Rome wasn't build in a day. As van der Linden (2006) himself did, the first step for a hierarchical approach is to find the appropriate models for the first level of modeling. Despite neither the Rasch model nor the log-normal model are breaking news in Psychometrics, their application to implicit measures data with the Linear Mixed-Effects model approach followed in this thesis is rather new. As such, we first wanted to find an appropriate and reliable approach to the separate modeling of implicit measures accuracy and time responses, able to be used as stand-alone models for each type of responses.

This thesis was mainly focused on the modeling of the IAT-family implicit measures, namely the IAT and the SC-IAT. Both the IAT and the SC-IAT are based on the accuracy and speed of the responses, and they exploit the logic of responses compatibility to sort different stimuli in contrasting conditions. The categorization happens by means of two response keys. Other implicit measures, such as the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001), exploits the same logic of response compatibility but in favor of the inhibition of the responses in contrasting conditions. As such, only one response key is needed. In the GNAT, only two categories at the time are presented, such as *Coke* and *Good*. Along with the stimuli belonging to these target categories, stimuli representing either other beverages or negative attributes are presented. The task is to identify the stimuli belonging to the target categories by pressing the response key and to do nothing (i.e., inhibit the response) when the distractors appear on the screen. The same task has to be performed in a contrasting condition where *Coke* exemplars and *Bad* attributes are the reference categories. The underlying idea is that it would be easier to press the response key when the reference categories are strongly associated between each other than when they are not. The structure and the type of task characterizing the GNAT make it not possible to obtain a response time of the correct inhibition when a distractor is presented. Consequently, the scoring of the GNAT is entirely based on the accuracy responses. Given that the accuracy and log-time response models presented in this thesis do not rely on each other to be applied, the model based on accuracy responses can be used to model the accuracy responses of the GNAT to obtain a Rasch parametrization of the data. If the GNAT is administered with other implicit measures, the accuracy responses of both measures can be modeled together with a comprehensive modeling such as that presented in this thesis.

So far, the modeling framework introduced in this work has been applied with main purpose of validating it, for both sand-alone implicit measures and multiple measures administered together. However, a more practical application is missing. For instance, this approach might be used for assessing the features of the IAT administration procedure on the respondents' performance. While it is known that some of features of the IAT administration, such as the order of presentation of the associative blocks, do influence the respondents' perfor-

mance (e.g., Greenwald, Nosek, & Banaji, 2003), the effect of other features, such as the presentation of a feedback, is less investigated. Following the LMMs approach of this thesis might be particularly useful for at least two reasons. First, it would allow to carry out the investigation in a latent variable modeling framework. Second, if the investigation of the effect of the administration features is carried out in a *within-subjects* experimental design, this approach allows for accounting for the dependencies of the observations. As such, it provides more reliable estimates, which in turn lead to more valid and generalizable inferences.

In the end

Despite the limitations, the results across the studies reported in this thesis highlighted the main aspects that have to be taken into account when analyzing IAT data.

Firstly, the consequences of not considering the fully-crossed structure of implicit measures and its related sources of variability and dependencies have been highlighted in terms of less reliable inferences of the constructs under investigation and a lower predictive ability of behavioral outcomes. The sensitivity of typical scoring methods to across-trials variability was blatant in the second empirical application of Chapter **??**. The predictive ability of the *D* score was improved just by reducing the across-trials variability with the selection of some of the stimuli. One would have expected that the performance of the *D* score computed on the least informative stimuli would have led to a worse prediction than both the one computed on the entire data set and that computed on the reduced data set containing only highly informative stimuli. Conversely, the reduction of the across-trials variability was the feature that mostly impaired the reliability of the *D* score. Even the *D* score computed on the least informative stimuli showed a better predictive performance than the one computed on the entire data set (i.e., the one mostly affected by the across-trials variability).

Another feature of interest is related to the use of differential measures. Across studies, differential measures showed their inadequacy for expressing the implicit construct under investigation. Differential measures resulted in a lower predictive ability than that provided by the linear combination of their single components.

Regardless of the methodology used for analyzing implicit measures data, the predictive ability of implicit measures was always outperformed by that of explicit measures. This result might be due to the fact that the behavioral task was presented right after the questions on the explicit chocolate evaluation. Consequently, the preferred chocolate might have been made salient by the explicit questions, and the choice might have been made accordingly. Another explanation can be given by considering the nature of the assessment provided by implicit measures. Indeed, implicit measures are supposed to measure the tendency to associate target objects, like the two types of chocolate, with positive and negative attitudes. Clearly, a measure like that reflects the like/dislike towards the specific target object. It is not a measure of how much one (or both) the target objects are wanted. According to Meissner et al. (2019), this is the feature of implicit measures leading to their low predictive ability of behavioral outcomes. Indeed, the choice might be more driven by a *wanting component* (i.e., how much an object is desired) rather than a *liking component* (i.e., how much an object is positively or negatively evaluated), which is the measure obtained from implicit measures. Nonetheless, the explicit assessment on the chocolate preference asked specifically how much respondent liked dark and milk chocolate. Consequently, also explicit measures aimed at the liking component and not at the wanting one.

Appendix A

R code for estimating Rasch model and log-normal model from IAT and SC-IAT data.

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

The example is based on the Coke-Pepsi IAT example of Chapter **??**. The estimation of the Rasch model and log-normal model from the SC-IAT data follows the same procedure. Consequently, the illustration is solely based on IAT data, but it can easily be implemented on SC-IAT data without any further changes besides the name of the data set.

This code can be copied and pasted in an R script, and it can be executed without changes as long as the data set on which the models are applied has the following characteristics:

- `subject`: Column containing the respondents' IDs (can be numeric, a factor, or a string, as long as it is unique for each respondent).
- `condition`: Column containing the labels for the two associative conditions of the IAT (SC-IAT) (factor with two levels such as `mappingA` and `mappingB`).
- `stimuli`: Column containing the labels identifying each stimulus (e.g., `good`, `bad`, `coke1`, `pepsi1`).
- `latency`: Column containing the latency of the IAT (SC-IAT) responses. Latency can

be expressed in seconds or milliseconds. In case the IAT (SC-IAT) included a built-in correction for the error responses, the raw response times should be used instead of the corrected ones.

- `correct`: Column containing the accuracy of the IAT (SC-IAT) responses, where 0 is the incorrect response and 1 is the correct response.

The data set must be in a long format. This means that the response of each respondent on each stimulus in each associative condition must be on a separate row, and the total number of observations (rows) for each subject must correspond to the total number of trials in the two associative conditions. For instance, in the IATs reported in Chapter **??**, respondents were presented with 60 trials in each associative condition, so that we had 120 trials for each respondent, and consequently 120 rows for each participant. In both the SC-IATs reported in Chapter **??**, respondents were presented with 72 trials in each associative condition, hence 144 observations (rows) for each respondent (in each SC-IAT) were obtained.

In both accuracy and log-time responses, the fixed intercept is set at 0, so that the estimates of the fixed effect of the IAT associative conditions can be interpreted as the expected *log-odds* of the probability of a correct response in each condition or the expected average log-response time in each condition, respectively.

For both accuracy and log-time responses, in Model 2 (Table **??** of Chapter **??**) the estimates of the stimuli are centered at 0 (argument `(1|stimuli)`), while in Model 3 (Table **??** of Chapter **??**) respondents estimates are centered at 0 (argument `(1|subject)`). In Model 1, the Null model, both stimuli and respondents are centered around 0.

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

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

A.1 Accuracy models specification

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

A.1.1 Model estimation

Model 1: The between–subjects variability is specified as random intercepts (i.e., `(1|subject)`). The between–stimuli variability is specified as random intercepts (i.e., `(1|stimuli)`) as well.

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

Model 2: The between–subjects variability is specified as random intercepts centered at 0 (i.e., `(1|subject)`). The within–stimuli between–conditions variability is specified as the random slopes of the stimuli in the conditions (i.e., `(0 + condition|stimuli)`).

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

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

```
a3 <- glmer(correct ~ 0 + condition + (1|stimuli) +
        (0 + condition|subject),
        data = data,
```

```
        family = "binomial")
summary(a3) # summary of the results
```

Model comparison

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

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

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

A.1.2 Rasch model parameters

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

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

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

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

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

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

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

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

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

Model 3 results in condition–specific respondents parameters and overall stimuli parameters. Respondents' condition–specific ability parameters can be extracted as follows:

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

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

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

A.2 Log-time models specification

The code for the estimation of the log-normal models is the same as the one used for the Rasch models. The changes concern the name of the specific function to use (from `glmer()` to `lmer()`) and the dependent variable (from `correct` to `log(latency)`). For this reason, only the code for the estimation of Model 3 and the code for extracting the log-normal model estimates is reported.

Model 3 can be estimated as follows:

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

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

A.2.1 Log-normal model parameters

We report the code for extracting the log-normal model parameters for log-time Model 3, assuming it was the best fitting model according to model comparison. The same code used for extracting the parameters for the accuracy models can be used for extracting the parameters of the log-normal models. The changes regard the name of the objects containing the models, from `a` to `t`, and the names of the new objects created for the parameters (e.g., from `easiness` to `intensity`).

Respondents' condition–specific parameters can be obtained as follows:

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

Stimuli overall time intensity parameters can be obtained as follows:

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

Appendix B

R code for a comprehensive modeling of implicit measures.

This appendix presents the R code used for obtaining the Rasch model and the log-normal model estimates from (Generalized) Linear Mixed-Effect models from IAT and SC-IAT data according to the comprehensive modeling approach presented in Chapter **??**

The example is based on the Coke-Pepsi IAT and the Coke SC-IAT presented in Chapter **??**. For illustration purposes, a Pepsi SC-IAT is considered as well. The associative conditions of the Pepsi SC-IAT are the Pepsi-Good/Bad one (PG condition) and the Pepsi-Bad/Good one (PB condition).

The data set should contain the following variables:

- `subject`: Column containing the respondents' IDs (can be numeric, a factor, or a string, as long as it is unique for each respondent).
- `measure`: Column containing the labels that identify the three implicit measures (e.g., `iat`, `cokesciat`, `pepsiciat`). This variable should be a factor with three levels.
- `condition`: Column containing the labels of the six associative conditions of the three implicit measures (e.g., `CGPB` and `PGCB` for the IAT, `CG` and `CB` for the Coke SC-IAT, `PG` and `PB` for the Pepsi SC-IAT). This variable should be a factor with six

levels.

- `stimuli`: Column containing the labels identifying each stimulus (e.g., `good`, `bad`, `coke1`, `pepsi`).
- `latency`: Column containing the latency of the IAT responses. Latency can be expressed in seconds or milliseconds.
- `correct`: Column containing the accuracy of the responses, where 0 is the incorrect response and 1 is the correct response.

The data set must be in a long format. This means that the response of each respondent on each stimulus in each associative condition of each implicit measure must be on a separate row, and the total number of observations (rows) for each subject must correspond to the total number of trials in the two associative conditions of each implicit measure. For instance, in the IAT reported in Chapter **??**, respondents were presented with 60 trials in each associative condition. Each of the SC-IATs was composed by 72 trials in each associative condition. The total number of observations (rows) for each respondent was 408 (i.e., 120 IAT observations, 144 Dark SC-IAT observations and 144 Milk SC-IAT observations).

Regardless of the dependent variable (i.e., either accuracy or log-time responses), the first model is the Null model in which both respondents and stimuli are specified as random intercepts across conditions and across implicit measures. The fixed effect is the effect of the implicit measure. Since the fixed intercept is set at 0, the estimates for each level of the fixed effect can be considered as the marginal *log-odds* (accuracy models) or the marginal expected average log-time response (log-time) models.

In the second model, the multidimensionality of the implicit measure is allowed at the respondents' level while stimuli are centered at 0. In other words, the random slopes of the respondents in the implicit measures (`0 + measure|subject`) and the random intercepts of the stimuli (`1|stimuli`) are specified.

Finally, in the third model the multidimensionality of the associative condition of the specific implicit measure is allowed at the respondents' level, by specifying their random

slopes in the associative conditions of each measure (`0 + condition|respondent`). Stimuli are specified as random intercepts (`1|stimuli`).

B.1 Accuracy models specification

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

B.1.1 Model estimation

Model 1: The effect of the implicit measure is specified as a fixed effect. The between–subjects variability is specified as random intercepts (`(1|subject)`). The between–stimuli variability is specified as random intercepts (`(1|stimuli)`).

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

Model 2: The effect of the implicit measure is specified as a fixed effect. The between–stimuli variability is specified as random intercepts centered at 0 (`(1|stimuli)`). The within–subjects between–measures variability is specified as the random slopes of the subjects in the implicit measures (`(0 + measure|subject)`).

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

Model 3: The between–stimuli variability is specified as random intercepts, centered at 0 (`(1|stimuli)`). The within–subjects between-conditions variability is specified as the ran-

dom slopes of the respondents in the conditions of each implicit measure (`(0 + condition|subject)`). The fixed effect is the associative condition of each implicit measure.

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

B.1.2 Model comparison

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

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

Models 2 and 3 have the same degrees of freedom. As such, the χ^2 statistics resulting from their comparison is meaningless, and only comparative fit indexes should be used instead.

B.1.3 Rasch model parameters

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

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

```
ability <- data.frame(
        subject = rownames(coef(a1)$subject), # Respondents' IDs
        ability = coef(a1)$subject[, 1] # Select only the random
)                       # intercepts estimates
```

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

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

```
)                          # intercepts estimates
```

Model 2 results in measure–specific respondents' parameters and overall stimuli parameters. Respondents' measure–specific ability parameters can be extracted as follows:

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

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

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

Model 3 results in condition–specific respondents ability parameters and overall stimuli easiness parameters. Respondents' condition–specific ability parameters can be extracted as follows:

```
cond_ability <- coef(a3)$subject[, -1] # drop the first column
                           # (fixed intercepts set at 0)
```

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

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

B.2 Log-time models specification

The code for the estimation of the log-time models is the same as the one used for the estimation of the accuracy models. The changes concern the name of the specific function to use (from `glmer()` to `lmer()`) and the dependent variable (from `correct` to `log(latency)`). Consistently, the code for extracting the log-normal model estimates

from the log-time models is the same as that used for extracting the Rasch model estimates from the accuracy models. For these reasons, only the code for the estimation of Model 3 and the related code for extracting the log-normal model estimates are reported.

Model 3 can be estimated as follows:

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

For the comparison between log-time models, the same code as the one used for accuracy models comparison can be employed. The names of the objects containing the models have to be changed accordingly, in this case from `a` to `t`.

B.2.1 Log-normal model parameters

The code for extracting the log-normal model parameters from log-time Model 3 is reported. The same code used for extracting the parameters for the accuracy models can be employed for extracting the parameters of the log-normal models. The changes regard the name of the objects containing the models, from `a` to `t`, and the names of the new objects created for the parameters (e.g., from `easiness` to `intensity`).

Respondents' condition–specific parameters can be obtained as follows:

```
cond_speed <- coef(t3)$subject[, -1] # drop the first column
                                     # (fixed interceptse set at 0)
```

Stimuli overall time intensity parameters can be obtained as follows:

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

References

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

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

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

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, *39*(12), 1–28.

Ellithorpe, M. E., Ewoldsen, D. R., & Velez, J. A. (2015). Preparation and analyses of implicit attitude measures: Challenges, pitfalls, and recommendations. *Communication Methods and Measures*, *9*(4), 233–252. doi: 10.1080/19312458.2015.1096330

Epifania, O. M., Anselmi, P., & Robusto, E. (2019). Dscoreapp: An user-friendly web application for computing the implicit association test d-score. *Journal of Open Source Software*, *4*(42), 1764. doi: 10.21105/joss.01764

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

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74–147. doi: 10.1080/10463280600681248

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

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

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

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

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

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

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

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

van der Linden, W. J. (2007). A hierarchical framework for modelling speed and accuracy. *Psychometrika*, *72*(3), 287–308. doi: 10.1007/s11336-006-1478-z

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

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