Chapter 1

Preface

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

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

The main objective of the Thesis is to provide new methods for more rigorous analyses of implicit measures data. In the long term, the repercussion of more rigorous data analyses can be observed in the replicability of the obtained results. For pursuing this aim, three paths are followed, one for a more sound approach to implicit measures data (sound path), one for a fairer comparison between implicit measures (fair path), and one for an easier (and more rigorous) way to compute implicit measures scores (easy path).

The sound path, which is also the main one, is an attempt at finding new approaches for the analysis of implicit measures. This is done by combining a classic of Psychometric Theories, the Rasch model, with a Linear Mixed Effects Models approach. The focus is mostly on one of the most popular, used, and studied implicit measures, the Implicit Association Test [**?**]IAT;¿Greenwald1998, and on its single category version, the Single Category IAT [**?**]SC-IAT;¿karpinski2006. Accuracy and time responses of implicit measures are modeled separately with distinct models. Consequently, the parameters are either explaining the processes leading to the accuracy responses or those leading to the time responses. The relationship linking these parameters can be explained and understood in a second level modeling [**?**].

Traditionally, Item Response Theory and Rasch modeling treats items (stimuli) as fixed factors (i.e., unknown constants that do not vary as a function of the observational units), while respondents are treated as random factors (i.e., effects that vary according to the observational units, drawn from a larger distribution) [**?**]. In this work, a slightly different approach was followed, also grounding on the data structure characterizing implicit measures. Indeed, the fully-crossed design characterizing the IAT (see Chapter **??**) allow one to conceptualize the stimuli as a manifestation of the super-ordered category they represent. Consequently, they are just one the possible set of stimuli that can be drawn from the same population of stimuli. Following this line of reasoning, it made more sense to consider them as random factors, and to treat them as random effects.

Besides being a statistically more sound approach, acknowledging the sampling variability of the stimuli implies that each stimulus potentially has a different functioning, and, consequently, a different impact on the observed responses. Therefore, if stimuli are treated as random and their random variability is accounted for, it is possible to exploit it for the best in order to gather all the information they convey [**?**]. Considering both respondents and stimuli as sample drawn from larger population and hence treating them as random factors allows for obtaining detailed and generalizable information not only at the respondents level, but also to consider, investigate, the functioning of each stimulus and its impact to the observed responses.

This approach can be followed also when multiple implicit measures are administered concurrently, which is a quite common practice in experimental psychology. In such cases, new sources of variability are added to the one already present in each implicit measure. By exploiting the flexibility of Linear Mixed Effects Models, a comprehensive modeling of multiple implicit measures within a Rasch approach is possible and can be used for gaining more reliable and comparable estimates at both the respondents and stimuli levels, for each implicit measure.

However, using Linear Mixed Effects Models for the conjoint analysis of multiple implicit measures within a Rasch framework is not a common approach to implicit measures data. The fair path is an attempt at providing scoring methods for a fairer comparison between the IAT and the SC-IAT in terms of their predictive capacity of behavioral outcomes. The approach followed in the fair

path is more in line with the typical analyses performed on implicit measures data. Size effect measures are the most popular and used scoring procedures for the IAT and the SC-IAT, and they are often used for comparing their performance on several variables used as criteria, such as the prediction of behavioral outcomes. The scoring procedures of both the IAT and the SC-IAT are affected by several artifacts, one of which is the lack of control on the sources of random variability in the data. Furthermore, both the scoring and the administration procedures present minor differences, such as the inclusion of a response time window or not, that might still influence the comparison in their predictive ability, ending up in misleading results. New scoring algorithms for the IAT and the SC-IAT are introduced in the attempt of minimizing not necessary procedural differences potentially affecting the comparison. The computation procedure of effect size measures itself cannot overcome the issues of the sources of random variability characterizing implicit measures data. By aligning the differences in the scoring procedure of the IAT and the SC-IAT, the new scoring alternatives should at least provide a means for a fairer comparison between the IAT and the SC-IAT. Consequently, the new, aligned, scoring algorithms should produce more reliable results regarding the comparison between the two measures on different criteria, such as the prediction of behavioral outcomes.

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

In the first Chapter, brief definitions of automatic and controlled processes are provided, and the main theoretical frameworks that have been proposed for conceptualizing the distinction between the two processes are outlined. The description of the IAT follows, along with the results of a literature review where the IAT use in different fields of application was investigated.

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

In these occurrences, the IAT is not able to provide the measure of interest. Other implicit measures have been introduced with the aim of providing an absolute measure of a target object. The SC-IAT [**?**] is a direct modification of the IAT procedure, where one of the target objects is dropped. It is often used as an alternative to the IAT when the aim is to obtain an absolute measure towards one object. The description of the SC-IAT is provided in Chapter **??**. The Chapter ends with a description of the fully-crossed design characterizing implicit measures, and with the reasons why this structure might undermine

the replicatibility of the results if it is not correctly accounted for.

Fair and easy paths are both presented in Chapter **??**. The typical and new, modified scoring procedures of the IAT and the SC-IAT are illustrated. The alignment of both administration and scoring procedures of the IAT and the SC-IAT in the new scoring algorithms, should provide a comparison between the predictive ability of the two measures more centered on the implicit measures themselves than on the scoring/administration avoidable differences. The comparison is usually based on the implicit measures predictive ability of behavioral outcomes, and the IAT tends to outperform the SC-IAT.

The rationale for the introduction of the new scoring algorithms is that, if by aligning the administration and scoring procedures of the two implicit measures, while acknowledging their main features, the IAT still outperforms the SC-IAT, its better performance should be ascribable to the measure itself, and not to procedural artifacts. The results of an empirical study where typical and modified scoring procedures were compared are reported. Regardless of the scoring algorithms, the measure obtained from the IAT always outperformed the one obtained from the SC-IAT. Limitations of these results might be related to the choice task considered as a behavioral outcome, as further illustrated in the Chapter.

Regarding the easy path component of Chapter **??**, two open source alternatives for the computation of the IAT and the SC-IAT typical scoring procedures are illustrated. One of them is a Shiny app (DscoreApp) for the computation of the IAT D score, while the other is an `R` package for the computation of the IAT and the SC-IAT D scores (`implicitMeasures`). DscoreApp was developed with the aim of providing researchers using the IAT an Open Source tools able to make the D score computation easier, without requiring for any programming experience. Beyond making the computation easier, it also guarantees for the replicability of the results. Indeed, researchers often fail to report the specific D score algorithm they have used for scoring the IAT. Additionally, replicability of the results is undermined by the high number of steps that are required for cleaning and preparing the data [**?**]. By automating the procedure and providing clear labels and descriptions for the identification of each scoring algorithm, these errors should be prevented, and the results replicability should be enhanced.

Among others, DscoreApp presents two main shortcomings. One of them is an intrinsic limitation of Shiny apps. By putting the code into the shiny interface, it is not possible to call it and run it from the command line, hence making it impossible to reproduce. However, this issue can be overcome by storing the code in a public repository, such as GitHub, as it was done for DscoreApp. Another important issue is that DscoreApp only computes the score for the IAT.

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

Chapter **??** provides an overview of the main modeling frameworks that have

been introduced for modeling IAT data. These frameworks can be distinguished according to the type of responses used for the estimation of the parameters. Quad model [**?**] and ReAL model [**?**] are based on accuracy responses, while Diffusion model [**?**] and Discrimination-Association model [**?**] account for both accuracy and time responses.

Regardless of the type of responses they consider, these models are able to disentangle the most automatic from the most controlled processes intervening during the performance at the IAT. A common finding is that automatic associations are just one of the possible processes intervening during the performance at the IAT, and that other controlled processes, such as recoding the stimuli (ReAL, Diffusion Model) or suppressing the automatically activated response (Quad model), play an important role as well. Despite their usefulness for the disentanglement of the IAT effect, these models come with some limitations. Most importantly, none of them can provide a detailed information at the level of the individual stimulus. This is a crucial point, also given that previous studies highlighted the importance of stimuli selection for a correct functioning of the IAT [**?**]e.g.,¿bluemke2006. Moreover, the fully-crossed structure of the IAT is overlooked.

Rasch modeling of IAT does provide a detailed information on the stimuli functioning. By pinpointing the stimuli that give the highest contribution to the IAT effect, it is possible to delve deeper on the automatic associations driving the IAT effect, and hence to have a better understanding of the measure itself. The applications of the Rasch model to IAT data that have been done so far are not save from criticisms. The most outstanding one is related to the discretization of the time responses, which might cause a large loss of information. Moreover, also Rasch modeling does not account for the random noise in the data. The random noise due to the different sources of variability in the IAT data brings sources of dependency that are very likely breaking the local independence assumption.

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

The log-normal model is introduced in the first section of Chapter **??** as well. By considering the normal density distribution of the log-time responses, the log-normal model allows for obtaining a parametrization of the data analogous to that provided by the Rasch model. Therefore, the discretization of the time responses needed for the application of the Many Facet Rasch Model (Chapter **??**) can be avoided. The estimates of the log-normal model parameters can

be obtained by applying Linear Mixed Effects Models (LMMs) to the IAT log-transformed time responses, hence the sources of dependency are addressed.

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

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

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

The second application was aimed at understanding whether the estimates provided by the proposed modeling framework do result in a better inference of the implicit construct under investigation, and, consequently, lead to a better prediction of behavioral outcomes, than the one given by the typical scoring procedure. The second application was also aimed at testing the usefulness of the condition–specific stimuli estimates. If the stimuli estimates truly allows for pinpointing the most informative, as well as the least informative stimuli, a higher amount of information should be obtained by selecting only the most informative stimuli. Therefore, a data set with a fewer number of stimuli, and hence, of trials is obtained. In such data set, the across trials variability due to heterogeneity of the stimuli should be reduced. Consequently, the D score

computed on the reduced data set should be more reliable than the one computed on the entire data set, and it potentially results in a better prediction of behavioral outcomes. An IAT for the implicit assessment of the preference for Dark or Milk chocolate (Chocolate IAT) was employed for pursuing these aims.

Rasch model and log-normal estimates did result in a better inference of the implicit preference that allowed for a better prediction of the behavioral outcome than the one provided by the typical scoring procedure. Moreover, the information on the contribution of each stimulus to the IAT effect allowed for reducing the across-trials variability, and the D scores computed on the reduced data set did result in a better prediction than those computed on the entire data set. Interestingly, even the D score computed on a reduced data set obtained by selecting only the least informative stimuli provided a better prediction than the one computed on the entire data set. Grounding on these results, it is possible to speculate that, no matter of the information provided by the stimuli, the across-trials is what mostly biases the D score computation, and just by reducing it, it is possible to obtain better measures more related to external variables.

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

The issues related to the data structure of implicit measures administered alone have already been highlighted in Chapter **??**. When multiple measures are administered concurrently, each of them comes with its peculiar data structure and its method variance. Other sources of dependency have to be expected, namely the within–respondents between–measures variability. Moreover, since usually different implicit measures employ the same set of stimuli, also the within–stimuli between–measures variability might be present. Therefore, on top of the method specific variance of each measures, also other sources of variability should be taken into account to obtain reliable estimates.

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

An empirical application of the modeling approach in Chapter **??** is pre-

sented in Chapter **??**. Data are the same as those in Chapter **??**, hence including one IAT and two SC-IATs. Implicit measures have been modeled separately with the (G)LMMS of Chapter **??** for obtaining Rasch model and log-normal estimates from each of them individually. This was done for mainly two reasons. Firstly, to investigate the soundness of the proposed approach for modeling other measures than the IAT. Secondly, to investigate whether and how model estimates change if the within–respondents between–measures variability and the within–stimuli between–measures variability is not accounted for. Results pointed out that, just by accounting for the method specific variance of each implicit measure, it is possible to obtain estimates more reliable than the D score, resulting in a better prediction of the choice. Nonetheless, by analyzing the data from each implicit measure separately, the estimates at the stimuli level might be misleading (e.g., it is not possible to rule whether the different functioning of the stimuli between measures is ascribable to an actual different functioning or to uncontrolled error variance). Moreover, the estimates at the respondents' level cannot be compared between implicit measures. The estimates obtained from the comprehensive modeling are similar to those obtained with the separate modeling of each measure. However, the comprehensive modeling allows for directly comparing the estimates at the levels of both respondents and stimuli. Consequently, a better understating of the functioning of each implicit measure is obtained, and more meaningful inferences can be made.

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