# Chapter 1

# Final remarks

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., ?, ?). 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, by distinguishing the most automatic components from the most controlled ones. 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. The influence of controlled processes needs to be taken into account when drawing conclusions based on IAT results, especially if the typical scoring method, the $D$ score, is used for expressing them. The $D$ score confound the contribution of automatically activated associations with that of other processes, such as the effort to overcome automatically activated bias (?, ?) or to use other strategies to simplify the task (?, ?, ?).

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. 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. Concerning the stimuli, the most fine-grained information provided by these models is at the stimuli categories level.

Given that the aim was to disentangle the respondent 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. blueResults in the literature 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. However, there are reasons to believe that the sources of variability and related dependencies go beyond the respondents, the stimuli, and the associative conditions.

blueDespite 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. However, some pieces of the puzzle were still missing. 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 observation, and with the possibility of being extended for modeling multiple 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. blueA less obvious and less straightforward feature of LMMs is their link with the Rasch model and the log-normal model, and specifically, how LMMs allows for estimating the parameters of these Psychometrics model. blueHowever, it must be considered that both the Rasch model and the log-normal model are

nothing else than linear models 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 defining 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 (?, ?). 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. Models with different random structures have been specified for analyzing IAT accuracy and log-time responses. 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. The feasibility of these models, their usefulness for the analysis of IAT data, and their comparison with typical IAT scoring 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, and hence 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 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., ?, ?).

This result is further corroborated by the low difference between the condition–specific estimates of the *Black* people category. 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 and 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, bluethe 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/Milk chocolate (Chocolate IAT). 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.

blueBy using dark and milk chocolate as target objects of the IAT, it was possible to reward the participation of the respondents to the experiment 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 either the differential measures and the linear combination of the their single components in different logistic regressions. Backward deletion and the model accuracy of prediction of the choice 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 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 (?, ?), 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 pools. 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. blueThis result might indicate that, despite respondents slowed down in one condition, their accuracy performance is still impaired in that condition. In a similar vein, this result might indicate that while respondents speeded up their performance in one condition, their accuracy performance is fostered 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 in one condition. 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. blueIn 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 the 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 difference between 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 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 were not able to 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. Only the speed in 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 contri-

bution 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 (?, ?). 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 other one 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. blueThe 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. blueA 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. blueThe 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 than the IAT is truly ascribable to the measure itself and not to some minor feature? By providing easy-to-compute and easy-to-interpret effect size measures with which typical users of these implicit measures might be more familiar with, might help in answering this question or, at the very least, in fostering a fairer comparison between the IAT and the SC-IAT.

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. In other words, the fair path was aimed at providing a more rigorous scoring methods for different implicit measures without moving apart from the typical approach.

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

blueThe 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 implicated in the performance at the IAT. Moreover, the information at the stimuli 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 construct under investigation.

blueHowever, 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 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.

blueThe 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 (?, ?). 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 (?, ?, ?). 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 ? (?). 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 the 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.

blueUndoubtedly, the second level of modeling introduced by ? (?) would provide further insights on implicit measures functioning, concerning both the stimuli and the respondents. Nonetheless, Rome wasn't build in a day. As ? (?) 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 indeed 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.

# In the end

*It doesn't even matter...*

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 on the construct 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 only 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 ?
(?), 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 eval-
uated), 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.