The Implicit Association Test and its Difficulty(ies): Introducing the Test Difficulty Concept to Increase the True-Score Variance and, Consequently, the Predictive Power of Implicit Association Tests

Merlin Urban, Tobias Koch, & Klaus Rothermund

Friedrich-Schiller-Universität Jena, Germany

Journal of Personality and Social Psychology, in press

Author Note

All study material, code, data and documentation of the AIID study is publicly available at the AIID's Open Science Framework page (Link: https://osf.io/pcjwf/). All other material, data, and code specific to our project, as well as the preregistrations of our specific hypotheses and analyses plans, are publicly available at our project's Open Science Framework page (Link: https://osf.io/ex9ar/). To date, only the narrative interpretations of Study 1 have been presented at the 64th Conference of Experimental Psychologists (TeaP), the narrative interpretations of all other research findings have not been publicly disseminated. We have no conflicts of interest to disclose.

Abstract

We introduce the test difficulty concept from classical test theory (CTT) to tackle the issue of low predictive power of Implicit Association Tests (IATs). Following CTT, we argue that IATs of moderate difficulty (defined as mean IAT scores of zero) have more predictive power than IATs of extreme difficulties (defined as mean IAT scores deviating strongly from zero). Furthermore, we assume this relationship to be mediated by the true-score variance in IAT scores, with moderate difficulty resulting in more true-score variance. To test our hypotheses, we used non-experimental (Study 1 and 2) and experimental designs (Study 3). In Studies 1 and 2, we compared IATs of different test difficulties with regard to their ability to predict direct attitude measures, drawing on the Attitudes, Identities, and Individual Differences (AIID) Study. In Study 1, a subset of 95 attitude IATs (n=127,259) was analyzed using multilevel structural equation models (SEM). As expected, IAT test difficulty strongly moderated the predictive power of IATs, and this effect was mediated by IAT true-score variance. In Study 2, we replicated the results with the same analyses but a different subset of 95 identity IATs (n=43,745). In Study 3, we experimentally manipulated IAT test difficulty. In total three IATs (n=480) were analyzed using multigroup SEMs. Again, the IAT closer to moderate difficulty had more true-score variance and predictive power than the IATs of extreme difficulty. Accordingly, for correlational research, we recommend developing moderately difficult IATs to maximize IAT true-score variance and provide suggestions on how to achieve that.

Keywords: Implicit Association Test (IAT), predictive power, classical test theory, test difficulty, individual differences

The Implicit Association Test and its Difficulty(ies): Introducing the Test Difficulty Concept to Increase the True-Score Variance and, Consequently, the Predictive Power of Implicit Association Tests

In recent years indirect measures are increasingly criticized for their low predictive power, that is, their poor ability to predict outcome variables (Blanton et al., 2009; Meissner et al., 2019; Oswald et al., 2013; Meissner & Rothermund, in press). Closely linked to this criticism is the Implicit Association Test (IAT; Greenwald et al., 1998), which is still the most prominent indirect measure and which is the indirect measure we will concentrate on in this article accordingly. In several meta-analyses, the IAT was found to be a weak predictor of behavioral measures with average correlations ranging from r = .14 to r = .27 as well as of behavioral measures holding direct measures constant with average correlations ranging from r = .14 to r = .18 (Greenwald et al., 2009; Kurdi et al., 2019; Oswald et al., 2013). In addition, meta-analyses also revealed that the IAT has low average correlations with its corresponding direct measures, ranging from r = .12 to r = .24 (Greenwald et al., 2009; Hofmann et al., 2005; Kurdi & Banaji, 2019; Oswald et al., 2013). Although some dual process theories (e.g., Strack & Deutsch, 2004) suggest not to expect high associations between the IAT and direct measures, most of the reported literature suggests a general increase in the predictive power of the IAT to be desirable.

Despite the fact that IAT correlations are low on average, the correlations are nevertheless heterogeneous. Accordingly, the effects of a number of potential moderators that could possibly explain the heterogeneous results were investigated (e.g., the correspondence between the IAT and the outcome variable, the type of the outcome variable, the type of the IAT, the IAT scoring method). Unfortunately, most moderators were found to have no or only mixed effects, leaving the problem of generally low predictive power of the IAT unresolved and also offering little indications for improving the IAT with regard to its predictive power (cf. Greenwald et al., 2009; Kurdi et al., 2019). As a consequence, further investigation of the

moderators of IAT relations is necessary and indeed ongoing (e.g., Gawronski et al., 2020; Irving & Smith, 2020).

In this article, we aim to contribute to this search by introducing a new perspective on the matter, taking a test theoretical approach. More specifically, we apply the test difficulty concept from classical test theory (CTT) to the IAT and relate it to and compare its potential utility with already examined test-theoretical concepts that have been applied to implicit measures. We then demonstrate the considerable influence of IAT test difficulty on the predictive power of IATs using both non-experimental as well as experimental research designs and discuss methodological as well as conceptual implications of our findings.

The difficulty concept and its influence on correlations

It is well known from CTT and questionnaire construction that the difficulty of an item or test is crucial for its ability to capture and predict individual differences and thus for its ability to relate to an outcome variable (e.g., Nunnally, 1970; Nunnally & Bernstein, 1994). Item or test difficulty is defined as answering an item or test in the keyed direction of the theoretical construct examined. The more people answer in the keyed direction, the easier the item or test and vice versa. The meaning of answering in the keyed direction of the theoretical construct depends on the theoretical construct examined. For example, in case of the attitude construct, to answer in the keyed direction means to give a response in favor of the attitude object. In accordance with this definition item or test difficulty is usually expressed as the mean of the item or the mean of the test (Nunnally, 1970; Nunnally & Bernstein, 1994). An item measuring the construct attitude is easy if most of the participants answer in favor of the attitude object, difficult if most participants answer in opposition to the attitude object and moderately difficult if some of the participants answer in favor and some in opposition or if they evaluate the attitude object as neutral. The same conceptual rationale of low, moderate, and high difficulty can also be applied to tests. While items and tests of moderate difficulty offer optimal conditions to predict individual differences, the same cannot be said about items

and tests of extremely low or high difficulty. This is because moderate difficulty typically allows for high interindividual variance, whereas extreme difficulties produce low variance (Gulliksen, 1945).

Applying the difficulty concept and its influence on correlations to IATs

The just described considerations can also be applied to the IAT if the IAT is considered a psychological test (e.g., Blanton et al., 2015; Fiedler et al., 2006). The test difficulty of an IAT can then be understood as the mean of the IAT, that is, as the average IAT effect. Due to the construction and the composition of the IAT (i.e., its relativity and block structure) two prerequisite conditions must be clarified in order to unambiguously interpret the mean IAT effect with regard to an IAT's test difficulty: 1) one of the two target categories must be selected as the target category in which keyed direction it is answered or not (we call that target category *relevant target category* from now on) and 2) the block in which the relevant target category and the attribute category expressing the keyed direction (e.g., the positive attribute category in case of an attitude IAT or the self-attribute category in case of an identity IAT) share the same response key must be defined. When defining this block to be the block that constitutes the subtrahend of the difference that produces the IAT effect, a positive (negative) average IAT effect indicates that the IAT is easy (difficult). An average IAT effect of (or close to) zero indicates that the IAT is moderately difficult.

We hypothesize that IAT test difficulty affects the predictive power of IATs as is known from CTT and questionnaire construction. Consequently, IATs of moderate difficulty should have more predictive power than IATs of low or high difficulty. This effect, in turn, should be mediated by differences in the variance of the respective IATs. Specifically, we hypothesize that IATs have stronger relations with outcome variables the closer their average IAT effects are to zero, and that IATs have weaker relations with outcome variables the further away their average effects are from zero. These differences in predictive power due to different levels of test difficulty should be explained by differences in variance in the IAT

scores of the respective IATs, with IATs of moderate difficulty having more variance compared to IATs of more extreme difficulties.

Although we are not the first to propose the described relationships between IAT effects, IAT variance and the ability of the IAT to predict outcome variables, the view on the matter has been quite controversial in the literature so far and empirical evidence is lacking.

Experimental vs. correlational approaches to research on implicit measures: Reliable effects vs. reliability of measures

Reviewing the literature on implicit measures in terms of whether large or small effect sizes, understood as (standardized) mean differences of the respective implicit measure such as the average IAT effect, are desirable for correlational research, one encounters conflicting views. Some authors argue that large effect sizes are better for studying individual differences (Perugini et al., 2010) or that they are even a prerequisite for studying individual differences (Kurdi & Banaji, 2017), while others point out that large effect sizes can be problematic for studying individual differences (De Schryver et al., 2016; Hedge et al., 2018). The reason for the conflicting views are first and foremost different ideas about whether large effect sizes are associated with greater variance in the true scores of the respective measures (Perugini et al., 2010) and thus leading to higher reliability coefficients (Kurdi & Banaji, 2017) or with lower variance in the true scores (De Schryver et al., 2016) and thus leading to lower reliability coefficients (Hedge et al., 2018). These divergent opinions about the relationship between effect size, variance, reliability, and correlation can be traced back to two historically different approaches in psychology, the experimental and the correlational approach (Cronbach, 1957).

In experimental psychology, researchers are interested in testing mean differences between experimental conditions. This goes along with experimental psychologists being interested in reducing variance or individual differences within experimental conditions (i.e. error variance for experimental psychologists) in order to increase effect sizes and the power of the test statistic with which experimental conditions are compared (Cronbach, 1957).

Variance within conditions is not only irrelevant for these comparisons in the sense that the amount of variance should not affect an existing mean difference in the population, an increase in variance even increases the standard error for testing these mean differences in samples. As variance within conditions represents individual differences the aim of reducing error variance to increase effect sizes of mean differences for the experimental psychologist goes along with reducing true-score variance of a measure. If the aim is to establish a "reliable" (i.e., robust and replicable) IAT effect that is different from zero (i.e., RTs that differ between the compatible and incompatible blocks of an IAT), then one should create an IAT which produces similar results for everyone, that is to say, an IAT with low true-score variance, since individual differences in the direction and/or strength of an IAT effect feed into the error variance of this test of mean differences.

In correlational research, the main focus is not on differences between experimental conditions, but on individual differences (Cronbach, 1957). The aim is to investigate associations between individual differences for different variables (predictors and outcomes). Consequently, the goal is not to have large effect sizes of mean differences between conditions, but to maximize the true-score variance of a measure. In this correlational context "reliability" is defined according to CTT as the proportion of true-score variance to observed variance in a measure (Lord & Novick, 1968). In case of the IAT, correlational research should thus try to maximize the true-score variance of the IAT measure.

Taken together, based on statistical theory, the claims that large effect sizes of mean differences such as the IAT D score are better or a prerequisite for correlational research (e.g., Kurdi & Banaji, 2017; Perugini et al., 2010) are unlikely to be true. This idea stems from experimental research, but neglects that as experimental researchers try to reduce individual differences to increase effect sizes of mean differences, large effect sizes are more likely to be associated with less true-score variance in a measure. Consequently, the claims that large effect sizes are associated with less true-score variance in a measure and thus

counterproductive for correlational research (e.g., De Schryver et al., 2016; Hedge et al., 2018) are more likely to be true. This is consistent with our hypotheses: Even when test difficulty is defined in the context of implicit measures via effect sizes, as in the case of the IAT and its D-score, extreme difficulties (i.e., extreme effect sizes) should be associated with lower true-score variance and consequently lower predictive power, whereas moderate difficulties (effect sizes around 0) should be associated with higher true-score variance and consequently higher predictive power.

However, it is important to note that the hypothesized relationships cannot be logically deduced from statistical theory, but are in need of empirical investigation. This is mainly because IATs of extreme difficulties (i.e., of large effect sizes) in comparison to most questionnaire measures of extreme difficulties (i.e., of large average scores) do not need to suffer from a restriction in range of possible values. In the case of IAT effects, the possible values of the distribution are virtually infinite due to the continuous response format of the IAT. Even when the D score algorithm is used to quantify the strength of an IAT effect, which ranges between values of -2 and +2 (Blanton et al., 2015), average IAT D scores typically do not even come close to these borders, and thus IAT distributions will not necessarily become more skewed when average IAT effects become larger or smaller (even with a large average IAT effect of D = 1 there is ample room for higher values, and the distribution must not be substantially skewed), but may even just shift from left to right or from right to left on the continuous scale. This is different compared to most questionnaire measures which are comprised of discrete response formats with a small and limited range (e.g., 7-point Likert scales) and which thus must become more skewed the larger their average scores (i.e., the more extreme their difficulties). The argument for why easy or difficult IATs should have less true-score variance and less predictive power compared to IATs with moderate difficulty is thus unrelated to range restrictions in possible values, and instead relates to the ability of the test to capture differences across the entire spectrum of possible

attribute values. Easy or difficult IATs should have less true-score variance because they only discriminate between extreme (and non-extreme) attribute characteristics, but unlike moderate difficult IATs, they are incapable of capturing differences at the center of the attribute distribution (e.g., medium and high values on the respective attitude attribute should both lead to similarly positive IAT scores for an easy IAT, only very low values on the attribute will produce a lower IAT score). In any event, it is an empirical question whether IATs of large effect sizes (very easy or difficult IATs) are associated with less true-score variance and thus with less predictive power.

As far as we know, the relations between effect size, true-score variance, as well as correlations with outcome variables have not yet been systematically investigated empirically in the context of implicit measures in general and the IAT in particular. Instead they have been simply assumed (Kurdi & Banaji, 2017), discussed theoretically (Perugini et al., 2010), or the focus of the empirical investigation was a different one (De Schryver et al., 2016; Hedge et al., 2018). By introducing the test difficulty concept to the IAT, we not only close this gap, we further help to structure the literature in a way that relates it to established concepts from psychometrics and test construction. While gaining a better understanding of so far unknown relations and enhancing theoretical soundness can be seen as worthy goals in themselves, the question arises as to why test difficulty is necessary and important to advance correlational IAT research, considering that the ultimate goal for correlational research is to increase the true-score variance of a measure. We will argue for the usefulness of the test difficulty concept in the following, explaining why we consider the concept of test difficulty to be the essential concept of this work.

The importance of IAT test difficulty for correlational research on the IAT

As described earlier the relationship between true-score variance of a measure and its predictive power is already well established in psychometrics. The challenge, however, lies in understanding how to effectively enhance the true-score variance of an IAT. Merely knowing

that true-score variance is important does not suffice in order to develop an IAT with substantial true-score variance, whereas understanding the relationship between test difficulty and true-score variance offers valuable insights into precisely that challenge. It aids in the development of IATs with increased true-score variance, surpassing what is known solely from the relationship between true-score variance and predictive power. To begin with, test difficulty in comparison to true-score variance arms researchers against common recommendations prevailing in the IAT literature that lead to a decrease in true-score variance. Furthermore, test difficulty is more easily intelligible and applicable in constructing IATs than true-score variance and as such, for example, offers guidance in assessing the suitability as well as in modifying the task design of an IAT for correlative purposes prior to conducting the actual empirical research.

Test difficulty – a potential safeguard against counterproductive recommendations for correlational IAT research

Understanding the relationship between test difficulty and true-score variance can help researchers to identify recommendations and research practices that are counterproductive for increasing the true-score variance of IATs and thus for correlative research. Given that our hypotheses are correct, one such recommendation stemming from the experimental origins of the IAT is to strive for large and robust IAT effects. When considering the entirety of the literature on implicit measures, rather than solely focusing on correlational research as we did previously, there is a majority of researchers advocating that large effects of implicit measures are desirable. In fact, large effects are widely regarded as a validation criterion for implicit measures per se (e.g., Greenwald et al., 1998; Greenwald et al., 2003; Payne et al., 2005). For example, the conventional scoring algorithm for computing the IAT effect was selected based on whether it maximized the effect size (Greenwald et al., 2003). Consequently, researchers commonly view striving for large IAT effects as a standard research practice, with the distinction between experimental and correlational aims often being overlooked. Even if

researchers make an effort to seek specific recommendations within correlational IAT research, they still find conflicting opinions about whether pursuing large and robust effects is beneficial or not as we described earlier. Due to this lack of awareness concerning the relationship between test difficulty and true-score variance, researchers may not realize that by striving for large and robust IAT effects, they most likely decrease the true-score variance of their IAT measures.

Another potentially counterproductive recommendation for increasing the true-score variance of IATs is to develop complementary IATs. IATs are complementary when the positive evaluation of one of the two target categories implies the negative evaluation of the other (e.g., pro-life vs. pro-choice) and they are non-complementary if the evaluation of both target categories is independent from one another (e.g., artists vs. musicians). Complementarity was one of the strongest and most robust moderators of IAT relations in the meta-analysis by Greenwald et al. (2009), with complementary IATs having more predictive power than non-complementary IATs. Consequently, researchers may use complementarity as a guideline to develop IATs with higher true-score variance and more predictive power. However, we will show that striving for complementarity might actually reduce true-score variance and predictive power under certain conditions, namely when complementarity leads to extreme IAT difficulties. Without understanding the relationship between test difficulty and true-score variance, there is no indication of when complementarity improves or decreases predictive power. At best, one would only know that complementarity generally leads to larger true-score variance and better predictive power, but sometimes it does not.

It becomes evident that understanding the relationship between true-score variance and predictive power alone is insufficient for developing IATs with increased true-score variance, as it does not help researchers to identify whether some of the most prominent and prevailing recommendations for the development of IATs, actually foster or hinder

correlational IAT research, and that understanding the relationship between test difficulty and true-score variance would add substantial value in this regard.

Test difficulty – a familiar terrain for IAT researchers and resulting positive consequences

True-score variance is defined in CTT as the variance of the conditional expectation of an observed variable as a function of the person variable (see Steyer, 1989) and is an unobserved, latent variable. Accordingly, complex analyses such as structural equation models are typically used for its estimation. Such analyses, however, are rather the exception in IAT research, among other things due to its experimental tradition. Consequently, the true-score variance of IATs is usually not reported and therefore remains unknown in most cases. Accordingly, most IAT researchers do not come into contact with the concept of true-score variance in their IAT research. We define IAT test difficulty, on the other hand, as the average IAT effect, which can be operationalized as a latent or a manifest variable. As the manifest average IAT effect is the most prevalent and important statistic in IAT research, test difficulty is nearly always estimated and reported regardless of the statistical analyses used. Therefore, IAT researchers are most likely more familiar with IAT test difficulty than with IAT true-score variance.

Immediate advantages to using the concept of test difficulty follow from this, for example, to assess the adequacy of an IAT for correlational purposes prior to conducting the actual empirical research. Arguably due to the described reasons researchers will be better able to estimate *a priori* whether an IAT will produce a large or small effect rather than whether an IAT will have more or less true-score variance. As such, IAT test difficulty compared to true-score variance can guide IAT researchers in assessing whether a particular IAT is well-suited for correlational purposes, given that IAT test difficulty is a valid proxy for true score-variance, saving time and resources and thus promoting correlational IAT research.

Test difficulty – its crucial yet overlooked role as a pivotal lever in influencing IAT true-score variance

While the concept of true-score variance is closely related to the predictive power of an IAT, it is not closely related to the development of an IAT. This is why so few factors influencing the true-score variance of IATs are known so far, none of which relate to the development, i.e., the construction of the task design, of the IAT. With the concept of test difficulty, it is the other way around. Test difficulty is not as closely related to the predictive power of an IAT as true-score variance itself, because its effect should be mediated by true-score variance, but compared to true-score variance, test difficulty is more closely related to the development of IATs and, thus, allows to deduce possible manipulations of the task design of IATs to increase IAT true-score variance. Not only does this provide more opportunities to influence the true-score variance and, consequently, the predictive power of IATs, but these opportunities can be particularly useful because the few factors that are known to influence true-score variance are typically not changeable for a given IAT, as we will show in the following.

One known influencing factor is the selection of the sample. By choosing a diverse sample relative to the phenomenon under study, IAT true-score variance can be amplified, as is already advocated in the literature (Greenwald et al., 2022). Another influencing factor is the selection of the to-be-measured construct. In case of attitudes, for example, there are attitudinal domains that may a priori be known to have inherently more true-score variance than other attitudinal domains (e.g., a Democrat/Republican attitude IAT is likely to have inherently more true-score variance than a Poor-People/Rich-People attitude IAT) or there are domains whose true-score variance depends on whether, for example, attitudes or self-concepts toward these domains are measured (e.g. an Environmental protection/Environmental degradation identity IAT might have more true-score variance than an Environmental protection/Environmental degradation attitude IAT). While knowing of these factors influencing IAT true-score variance is essential, it quickly becomes apparent that their practical utility can be limited. Often, researchers are interested either in the general

population or in a subpopulation with specific characteristics (e.g., people diagnosed with depression), or they are bound to a specific sample due to limited resources (e.g., college students). In these cases, choosing a different sample that is more diverse is not an option. Furthermore, researchers are typically interested in investigating a particular construct which might be known to have inherently low true-score variance (e.g., attitudes towards a specific minority); again, in this case measuring a different construct with the IAT altogether, for which more attitudinal true-score variance is to be expected, is not an option because it does not match the research question. In other words, although factors influencing IAT true-score variance are known, their number is limited and they remain immutable in many scenarios, thereby leaving researchers perplexed and uncertain about how to enhance IAT true-score variance and in turn the predictive power of IATs.

In situations like these, that is, when the to-be-measured construct (e.g., attitudes vs. identity), and the to-be-investigated population are already given by the research question and are thus immutable, IAT test difficulty comes into play. IAT test difficulty opens up new avenues for influencing IAT true-score variance, which center around the task design of the IAT. As such, they are independent of the hitherto established influencing factors and can be used even in situations where these established influencing factors are given. Drawing on the concept of test difficulty, we derive strategies for modifying the reference category, attribute categories, and exemplar stimuli in order to shift IATs towards moderate difficulty, thereby amplifying IAT true-score variance and bolstering the predictive power of IATs.[1] Of these strategies we consider the modification of the reference category to be the most promising

[1] Note that IAT test difficulty is not only influenced by the newly proposed design-related approaches, but also by factors that have been described as influencing IAT true-score variance, namely the sample and the to-be-measured construct. The dependency of test difficulty on the sample is well known (e.g., Hambleton & Jones, 1993; Lord, 1953), and just as some constructs have inherently more or less true-score variance, some constructs inherently tend to result in IATs of extreme or moderate difficulty (e.g., the Democrat/Republican attitude IAT is more likely to be of moderate difficulty whereas the Poor-People/Rich-People attitude IAT is more likely to be easy or difficult depending on the definition of the relevant target category). Another influencing factor is the context in which the IAT is administered; however, its influence on IAT test difficulty and IAT true-score variance is more complex (see the general discussion for possible influences of the context on IAT test difficulty and true-score variance).

one. Accordingly, we investigated this approach more closely (see Study 3 for theoretical explanations and empirical results) while we will address the other approaches in the general discussion. Note that the concept of test difficulty was a prerequisite for the development of the proposed strategies, even though we assume that its effect on the predictive power is mediated by true-score variance.

Taken together, we understand the concept of test difficulty as a pivotal lever that helps researchers to address the challenge of low IAT true-score variance and predictive power.

Hypotheses and methodological approach

The above arguments regarding the relation between test difficulty, true-score variance, and predictive power of IATs can be translated into the following research hypotheses:

1) Moderation hypothesis: The test difficulty of an IAT moderates its predictive power in such a way that IATs with more extreme (high or low) difficulty have weaker associations with outcome variables than IATs with moderate difficulty.
2) Mediated moderation hypothesis: This moderating effect of IAT test difficulty can be explained by the corresponding true-score variance of the IAT, that is, the moderating effect is mediated by the true-score variance of an IAT, with IATs of moderate difficulty having more true-score variance than IATs of extreme difficulty.

We investigated our hypotheses by using direct attitude measures as outcome variables, primarily because they are comparable across different IATs and, accordingly, eliminate undesirable influencing factors that otherwise could not be held constant, thus providing a good starting point for demonstrating the potential of the test difficulty concept to increase the predictive power of IATs.

The remainder of the article's structure follows from our arguments for why test difficulty has additional value beyond true-score variance for correlational IAT research. Our

initial goal was to establish the IAT test difficulty account and thereby to provide initial evidence of its usefulness, since with its establishment it would become clear that existing recommendations hinder correlational IAT research and that IAT test difficulty is a valid and viable proxy for IAT true-score variance. To this end, we used a non-experimental approach without manipulating IAT test difficulty, with which we tested a large number of IATs of varying difficulty and corresponding direct attitude measures in terms of our hypotheses by resorting to the Attitudes, Identities, and Individual Differences (AIID) Study (Hussey et al., 2018). The AIID Study is an extensive online study that ran on Project Implicit between 2004 and 2007. Our proceeding in this non-experimental approach was twofold. In a first step, we investigated a subset of the AIID study consisting exclusively of attitude IATs (hereafter referred to as Study 1), and in a second step, we investigated a subset of the AIID study consisting exclusively of identity IATs (hereafter referred to as Study 2). In doing so, we replicated our results from Study 1 with different to-be-measured constructs as well as with different samples and contexts in Study 2. We then sought to provide further evidence for the validity of the test difficulty account while at the same time providing further evidence for its usefulness by showing that IAT test difficulty can be used to derive design-related modifications to the IAT to increase the true-score variance and in turn the predictive power of IATs. To this end, we switched from a non-experimental to an experimental approach. In this final study (hereafter referred to as Study 3), we tested our hypotheses by manipulating IAT test difficulty via modifying the reference categories for an environmental attitude IAT with a fixed relevant target category (environmental protection).[2]

[2] Note that in all of the following analyses, we differentiate between total variance (i.e. manifest meta-analytical results of Studies 1 and 2) and true-score variance (multilevel SEM results of Studies 1 and 2 and multigroup SEM results of Study 3). For simplicity, we use the term "true-score variance" in both the multilevel SEM and the multigroup SEM to refer to the error-free or error-corrected variance in IAT scores. Note, however, that compared to the multigroup SEM, we used only one indicator for the IAT scores in the multilevel models, and therefore the definition of “true-score variance” in CTT as the variance of the conditional expectation of an observed variable given the person variable (Steyer, 1989) is not perfectly met in these models. Nevertheless, in the multilevel SEM as well as in the multigroup SEM the variance of the IAT scores is a latent variable that was corrected for unsystematic error variance, and accordingly it is still appropriate to refer to it as “true-score variance”.

Study 1

The aim of Study 1 was to examine whether the application of test difficulty to the IAT leads to results that are consistent with what CTT predicts. Therefore, we investigated whether IAT test difficulty moderates the relationship between IATs and outcome variables (H1) and whether this effect is mediated by IAT true-score variance (H2) in the proposed ways. To this end, we drew on a subset of the AIID study that consisted entirely of attitude IATs and which provided us with a large number of attitude IATs of varying test difficulty and corresponding direct attitude measures.

Methods

Design and procedure of the AIID study

In a first step of the AIID study, participants were asked to create a project implicit user ID and to answer demographic related questions. They then were randomly assigned to 1 out of 95 different domains, i.e. 1 out of 95 different target category pairs such as "Democrats" vs. "Republicans".[3] After that they had to complete one IAT and 29-33 self-report items assessing explicit attitudes for the particular domain. Two IAT-types were employed, an attitude or an identity IAT.[4] The ratio of attitude and identity IATs was 3:1 for each domain. This ratio equals the number of different attribute category pairs used. In case of the identity IAT the attribute category pair was self/other and in case of the attitude IAT the attribute category pairs were positive/negative, good/bad, or pleasant/unpleasant. The four attribute category pairs were randomized within each domain. In addition the attitude IATs in comparison to the identity IATs also varied with regard to the set of attribute stimuli used. The attribute stimulus sets were identical for an attitude IAT in a given domain, but were randomly assigned between domains with seven different sets of attribute stimuli in total. The

[3] Note that we assume the domains to be drawn at random from a distribution of possible domains and that we therefore assume the domains to be interchangeable.

[4] For Study 1, we are focusing only on the attitude IATs of the AIID study, whereas in Study 2 below we test our hypotheses on the basis of the identity IATs.

self-report attitude items were drawn randomly with constraints from a pool of a total of 93 possible attitude items. The order of the IAT and the self-report measures was randomized. Due to the fact that participants had user IDs they could participate in the study repeatedly. In this case they were assigned to a new domain without replacement. For more information on the study procedure, please see the Open Science Framework (OSF) page of the AIID study (Link: https://osf.io/pcjwf/).

Sampling and observations

For the purpose of our research we differentiated between IATs as observations and participants as observations. With respect to the IATs, the study consisted of 95 observations, one attitude IAT for each of the 95 domains. With respect to the participants, the study consisted of 137,502 observations in total (exploratory as well as confirmatory dataset of the AIID study). After excluding participants to ensure data quality (see results section for further details), we were left with an overall sample of 127,259 participants. The number of participants was relatively evenly distributed among the different IATs ranging from 944 to 1,723 with an average of 1,340 participants per IAT (*SD* = 164.3). The sample was rather diverse in terms of demographic criteria: from those who gave information on the corresponding demographic data 57.9% were 30 or younger, 40% were between 31 and 60, and 2.1% were over 60; 65.6% were female and 34.4% were male; 77.7% were from the United States, and of the 22.3% non-United States respondents about half came from Australia, Britain or Canada; 54.1% had a university degree, 34.8% had a college or associate's degree, and 11.1% reported having a high school diploma or less education.

Measures

Indirect Measures: IATs. The standard IAT block structure by Greenwald et al. (2003) was used for all 95 attitude IATs. The block order, i.e., which of the two combined tasks was presented first, was randomized for all IATs. The number of attribute stimuli per attribute category was 6 and all attribute stimuli were always words. In contrast, the target

stimuli varied in number and in format. Depending on the IAT the number of target stimuli varied from three to nine stimuli and the format varied between words, pictures, and a mixture of both. Due to the fact that the number of trials in the blocks were the same for all IATs, the number of presentations per stimulus varied. Attribute and target stimuli were presented in alternating order, but the order within the attribute and target stimuli themselves was randomized. For more information on the single attitude IATs, see the OSF page of the AIID study again (Link: https://osf.io/pcjwf/).

IAT effects were computed based on the D score algorithm (Greenwald et al., 2003). We recoded the data by exchanging the target categories X and Y when the mean D scores were negative so that the mean D scores of all IATs ranged from 0 to 2 (and not from -2 to 2) with positive IAT scores indicating a preference of target category Y over X. In this case D scores of 0 indicate the IAT to be of medium difficulty as already described in the theory section, while D scores larger than 0 indicate that the IAT is easy/difficult, with an increasing extremity the more the D score deviates from zero. Whether the IATs are easy or difficult depends on which target category is defined as the relevant target category (consequently, one and the same IAT can be easy or difficult). However, a distinction between easy or difficult IATs is irrelevant for answering our hypotheses, since we want to determine whether moderate IATs have more true-score variance and make better predictions than both easy and difficult IATs. Recoding the D scores in the manner described ensures that linear analyses are appropriate for examining the relationships between IAT test difficulty, IAT true-score variance and IAT relations.

Direct Measures: Gut Reactions, Actual Feelings, and Preferences. The constraints of the self-reported attitude items in the AIID study were such that only 5 of the 93 items were presented to every participant. In order to guarantee the largest possible sample size while at the same time reducing the number of missing values to a minimum, we focused on these 5 items. Two of the items, one for each target category, asked about the gut reactions

towards the target categories, two of the items, one for each target category, asked about the actual feelings towards the target categories, and the final item asked about which of the two target categories one preferred. The wording of the items was: a) "Rate your gut reactions towards category X, b) "Rate your gut reactions towards category Y", c) "Rate your actual feelings towards category X", d) "Rate your actual feelings towards category Y", and e) "Which do you prefer, category Y or X?" (X and Y were replaced by the target categories; e.g., X = Republicans and Y = Democrats). Items a) to d) were to be rated on a 10-point bipolar scale with endpoints ranging from 1 *(strongly negative)* to 10 *(strongly positive)*. Item e) was to be rated on a 7-point scale with endpoints ranging from 1 (*strongly prefer X to Y*) to 7 (*strongly prefer Y to X*). We calculated difference scores for each participant between the ratings of the two target categories, once for gut feelings and once for actual feelings, and recoded the item that asked about the preference so that all scores ranged from -9 to 9. As a final outcome variable, we then averaged all scores, with positive scores indicating a preference for target category Y over X. Difference scores for the direct measures were calculated in accordance with the difference that also defined the corresponding IAT scores.

Data analysis

The study design implies a multilevel data structure in which the individual scores are nested within the 95 different domains. Furthermore, participants were allowed to take part in the study repeatedly, in which case the same participant was assigned to a different domain, leading to additional dependency in the data. Although theoretically possible, we decided against reporting cross-classified multilevel models for several reasons,[5] opting instead for

[5] First, the user ID showed a substantial number of missing values, which in case of fitting cross-classified models results in dropouts (i.e., loss of information and statistical power). Second, the variances of the direct attitude scores and the IAT scores on a person-level were close to zero, which made it difficult to estimate these variances and to ensure proper convergence of the Monte Carlo Markov Chains (without the use of informative priors). Third, when nevertheless modeling cross-classified multilevel models to test the hypotheses, the results did not differ substantially from the reported multilevel models, at least with respect to our first hypothesis (see our OSF page for the results of the cross-classified multilevel moderation model). In the case of our second hypothesis, the model showed convergence problems, suggesting the need of model simplification, which was achieved by using a traditional multilevel model in the main analyses.

traditional multilevel models. In the following we outline how both our hypotheses translate into multilevel models.

Moderation hypothesis (H1). According to our first hypothesis IAT test difficulty moderates the strength of the relationship between IATs and direct attitude measures in such a way that the relationship increases the closer an IAT is to medium difficulty. For a general overview, Figure 1 Panel A shows a conceptual translation of this hypothesis into a multilevel model. Below, we will derive the proposed model step by step. First, we assume that the relationship between individual IAT scores and individual direct attitude scores varies across clusters, i.e., the 95 attitude domains. Both these variables consist of *within cluster variance* and are thus Level 1 variables (for a more detailed distinction between within and between cluster variance see Asparouhov & Muthén, 2006). Accordingly, we model a random slope at Level 1 (note for the sake of model completeness that we also model a random intercept at Level 1). We further assume that the varying relationship across clusters, that is, the random slope, can be explained by the moderator IAT test difficulty, that is, the cluster mean scores of the IATs. Both these variables only consist of *between cluster variance* and are thus Level 2 variables. Therefore, we additionally model a *Between effect* (c.f. Preacher et al., 2016) of the Level 2 variable IAT test difficulty on the Level 2 random slope variable, which is equivalent to a *cross-level interaction* of the Level 2 variable IAT test difficulty and the Level 1 variable individual IAT score. Finally, as is common practice in moderation analyses, we also model the main effect of the moderator IAT test difficulty (e.g., Hayes, 2013). This main effect equals the Between effect of IAT test difficulty on the between cluster variance of the direct attitude measure, that is, the cluster mean scores of the direct attitude measure, because variables consisting of between cluster variance can merely affect other variables consisting of between cluster variance (Curran & Bauer, 2011; Preacher et al., 2010).

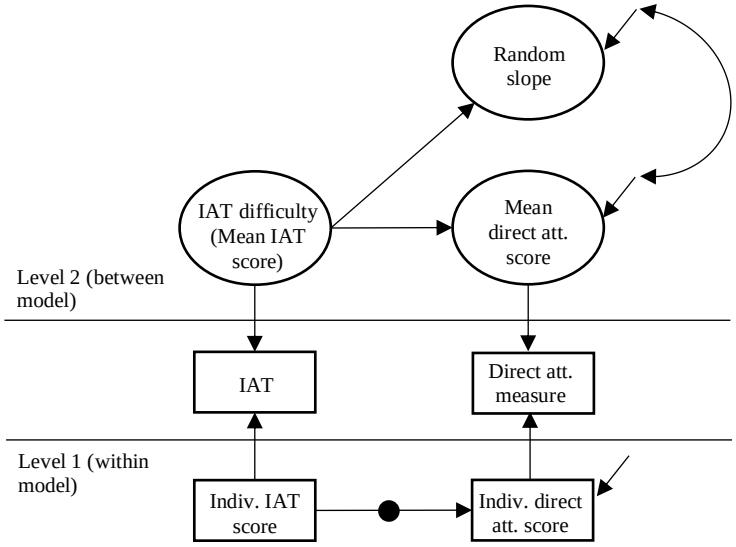
In order to circumvent problems of the standard multilevel paradigm, such as modeling observed cluster means or the conflation of within and between variance, we turned

to Preacher et al.'s (2016) proposed *multilevel structural equation modelling* (MSEM) paradigm. Consequently, we estimated latent cluster means following Lüdtke et al. (2008) and centered the individual IAT scores at these latent cluster means, which can be considered as a form of centering within clusters (CWC; see Enders & Tofighi, 2007). Within the MSEM paradigm, we used the *random coefficient prediction* (RCP) approach, also known as *slopes as outcomes* approach, to test the proposed multilevel model (Preacher et al., 2016). For a more formal description of our proposed multilevel moderation model see the online Supplement 1 on our OSF page (Urban et al., 2024; Link: https://osf.io/ex9ar/).

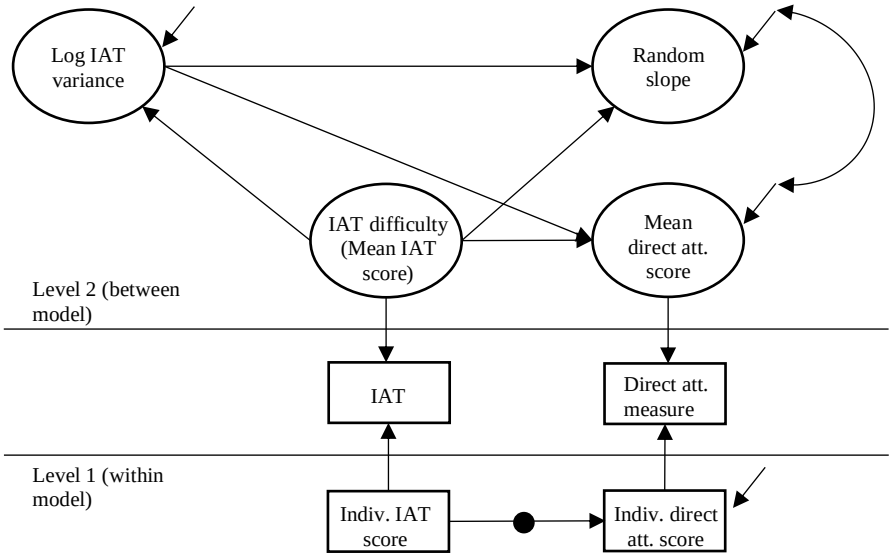
Figure 1

Conceptual Representation of the Multilevel Moderation Model (Panel A) and the Multilevel Mediated Moderation Model (Panel B)

A



B



Note. Circles represent latent variables and rectangles observed variables. Single-headed arrows are path coefficients, double-headed arrows are correlation coefficients, and unattached arrows are residuals. Darkened circles represent random slopes, which are also depicted on Level 2. IAT = Implicit Association Test; Log IAT variance = Log IAT true-score variance; att = attitude; Indiv = individual.

Mediated moderation hypothesis (H2). According to our second hypothesis, IAT true-score variance mediates the moderating effect of IAT test difficulty on the relationship

between IATs and direct attitude measures in such a way that the positive influence of medium difficulty on the relationship is due to an increase in true-score variance. Figure 1 Panel B shows a conceptual translation of this hypothesis into a multilevel model. In what follows, we will derive only the newly proposed part of the model. We postulate a mediation model, i.e., we assume an indirect effect of IAT test difficulty on the random slope, once IAT true-score variance is included in the model as a mediator. IAT true-score variance only consists of between cluster variance and is thus a Level 2 variable (note that we estimated the log variance to ensure a linear relationship between the variables). Hence, we model a Between effect of IAT test difficulty on log IAT true-score variance and a Between effect of log IAT true-score variance on the random slope, the latter being equivalent to a cross-level interaction of the Level 2 variable log IAT true-score variance and the Level 1 variable individual IAT score. In a final step, we model the main effect of log IAT true-score variance, which corresponds to the Between effect of log IAT true-score variance on the cluster mean scores of the direct attitude measure.

As in the case of our first hypothesis, we used the RCP approach within the MSEM paradigm to test the proposed multilevel model. A more formal description of the model can again be found in Supplement 1.

Transparency and Openness

The AIID study received research ethics committee approval from the University of Virginia. All material of the study including study material, raw data, data cleaning code and the curated exploratory and confirmatory dataset is publicly available on the AIID's OSF page (Hussey et al., 2018; Link: https://osf.io/pcjwf/). The exploratory and confirmatory dataset were the starting point for all further analyses specific to our project, for which we also provide all code on OSF (Urban et al., 2024; Link: https://osf.io/ex9ar/). The estimation of the multilevel models was carried out in Mplus (version, 8.4, Muthén & Muthén, 1998-2017), which allows researchers to automatically compute latent cluster means as well as the

log variance. R (version, 4.0.2, R Core Team, 2021) was used for all other analyses, such as data preparation for the multilevel analyses. Information on data exclusion and statistical power can be found in the results section. The specific hypotheses and analysis plans for our project were developed using the exploratory dataset, preregistered, and then tested using the confirmatory dataset, allowing for sound theoretical conclusions. The preregistration is also publicly available on OSF (Urban et al., 2024; Link: https://osf.io/ex9ar/).

Results

Preliminary analyses

Ensuring data quality. For large datasets generated online it is of particular importance to ensure data quality. To achieve this goal we applied several criteria. Trials were excluded according to the calculation of the D score (Greenwald et al., 2003). Participants were excluded when they did not have complete IAT data or when their responses indicated that they did not work properly on the IAT, following recommendations by Hussey et al. (2018). The exact criteria involved in Hussey et al.'s (2018) recommendations are described in more detail in Supplement 2.

Assessing the need for multilevel models. We tested the amount of dependency in the data by calculating the intraclass correlations (ICCs). The ICC of both, attitude IAT scores and direct attitude scores, was substantial, $ICC_{\text{attitude IAT}} = 0.20$ and $ICC_{\text{direct attitude}} = 0.19$, respectively, suggesting a considerable dependency in the data. Furthermore, we examined the variance of the random coefficients in the unconditional random intercept and random slope model (i.e., model without Level 2 predictors). The variance of the random intercepts and the random slopes was substantial as well, $\widehat{\sigma}_I^2 = 2.92$, with a 95% credible interval of *C.I.* [2.18, 3.93], and $\widehat{\sigma}_S^2 = 2.07$, with a 95% credible interval of *C.I.* [1.55, 2.81], respectively. Again, the result suggests strong variation in the data across clusters. Furthermore, both results also indicate that Level 2 predictors should be considered (Geiser, 2011), which is in line with our proposed models.

Bayesian analysis, handling missing values, and descriptive statistics. We used Bayesian estimation and Markov Chain Monte Carlo methods implemented in Mplus to analyze the data (see Asparouhov & Muthén, 2010, for details). Bayesian methods have some advantages over classical maximum likelihood estimation by allowing researchers to incorporate prior information (Depaoli & Clifton, 2015; Lee & Song, 2004), computing credibility intervals for all parameters in the model (Koch et al., 2016), and facilitating the estimation of complex multilevel models (Asparouhov & Muthén, 2010). Several studies have also shown that Bayesian methods outperform maximum likelihood estimation methods for single and multilevel models in terms of higher convergence rates, fewer improper solutions, and a more efficient and stable estimation process, especially in small samples (Depaoli & Clifton, 2015; Lee & Song, 2004; Zitzmann et al., 2015). In the present study, we used Bayesian methods and uninformative (default) priors implemented in Mplus because there was no sufficient information for the specification of informative priors and the data set was relatively large. Another advantage of Bayesian estimation is that it inherently handles missing values if the missing data mechanism is ignorable (see Asparouhov & Muthén, 2010; Gelman et al., 2014). In the present study the independent variable individual attitude IAT scores had no missing values due to the criteria applied to ensure data quality, but the dependent variable individual direct attitude scores had 6.62% missing values.[6] We used Bayes estimation to handle these missing values accordingly. Descriptive statistics of the observed variables can be found in Supplement 3.

Main analyses

Moderation hypothesis (H1). Standardized results of the full multilevel moderation model are displayed in Figure 2 Panel A (the unstandardized results can be found in

[6] Note that only cases where both difference scores could not be calculated due to missing values on the single items and where the preference item was not rated either were given missing values on the final score, while cases where at least one difference score could be calculated or the preference item was rated were entered into the analyses as aggregated difference scores.

Supplement 3). The standardized within effect of attitude IAT scores on direct attitude scores averaged across all domains was significant, with b = .40, a posterior standard deviation of pSD = .003, and a 95% credible interval of *C.I.* [.40, .41]. Overall, 18% of the variance in direct attitude scores could be explained by attitude IAT scores, *C.I.* [17.6%, 18.3%]. That is, participants with higher attitude IAT scores have also higher direct attitude scores on average. The hypothesized between effect of IAT test difficulty on the random slope, i.e., the cross level interaction of IAT test difficulty and attitude IAT scores, was also significant, with a standardized regression coefficient of b = -.35, pSD = .09, 95% *C.I.* [-.52, -.16], suggesting a moderating effect of IAT test difficulty. As expected the positive within effect of attitude IAT scores on direct attitude scores is stronger for IATs the closer their mean D score is to zero (IATs of medium difficulty) and weaker for IATs the further away their mean D score is from zero (IATs of extreme difficulty). Figure 3 illustrates this relationship (see Supplement 3 for a list of the estimates by domain).[7] IAT test difficulty explained 12.4% of variation in the slopes, *C.I.* [2.4%, 27.4%], using the Mplus R^2 option.[8]

The reported amount of explained variance R^2 of the above analysis is not comparable on a one-to-one basis with the results of previous meta-analyses in which moderators of IAT relations were tested. This is due to statistical and technical differences between our

[7] Note that the figure is based on unstandardized estimates since Mplus does not provide the corresponding standardized estimates.

[8] We ran the multilevel moderation analysis again including additional moderators on Level 2 that already have been identified to be promising moderators in previous meta-analyses (Greenwald et al., 2009; Kurdi et al., 2019). As such we included *complementarity* and *social sensitivity*. Other previously identified moderators, for example, *correspondence* or *type of the IAT* were constant across domains and thus could not be examined. The inclusion of additional moderators did not change the central results. The hypothesized between effect of IAT test difficulty on the random slope remained significant, with a standardized regression coefficient of b = -.42, pSD = .09, 95% *C.I.* [-.58, -.25]. In total the moderators explained 30.9% of variation in the slopes, *C.I.* [15.2%, 46.6%]. A comparison of the moderation model that included all three moderators with the moderation model that included only the two additional moderators, complementarity and social sensitivity, revealed that IAT test difficulty contributed 17.8% to the 30.9% of explained variation in slopes. Detailed descriptions of how the moderators were selected can be found in the general discussion, and detailed descriptions of how they were coded as well as of the results of the corresponding models can be found in Supplement 3. We also conducted additional analyses in which we controlled for moderating effects of the true-score variance in direct attitudes on the random slope as a potential confounding variable. The results of these analyses are reported in Supplement 3. Importantly, test difficulty predicted the random slope over and above the true-score variance in the direct attitude measures.

multilevel approach and typically used approaches (e.g., meta-analytical approaches). These differences mainly stem from different formulas to estimate R^2 in multilevel models using Mplus vs. R^2 in meta-analytical models using other statistical programs (e.g., Nakagawa & Schielzeth, 2013; Snijders & Bosker, 2012), but also from using latent cluster means (instead of manifest cluster means), estimating random regression coefficients (instead of correlations), using Bayes (instead of frequentist maximum likelihood) estimation, and using Bayes to handle missing values (instead of listwise deletion). To better compare our results to previous meta-analytical results, we also conducted a meta-analysis. In this case IAT test difficulty explained 24.1% of the variance in correlations between attitude IAT scores and direct attitude scores across domains (a complete description of the meta-analytical results can be found in Supplement 3).

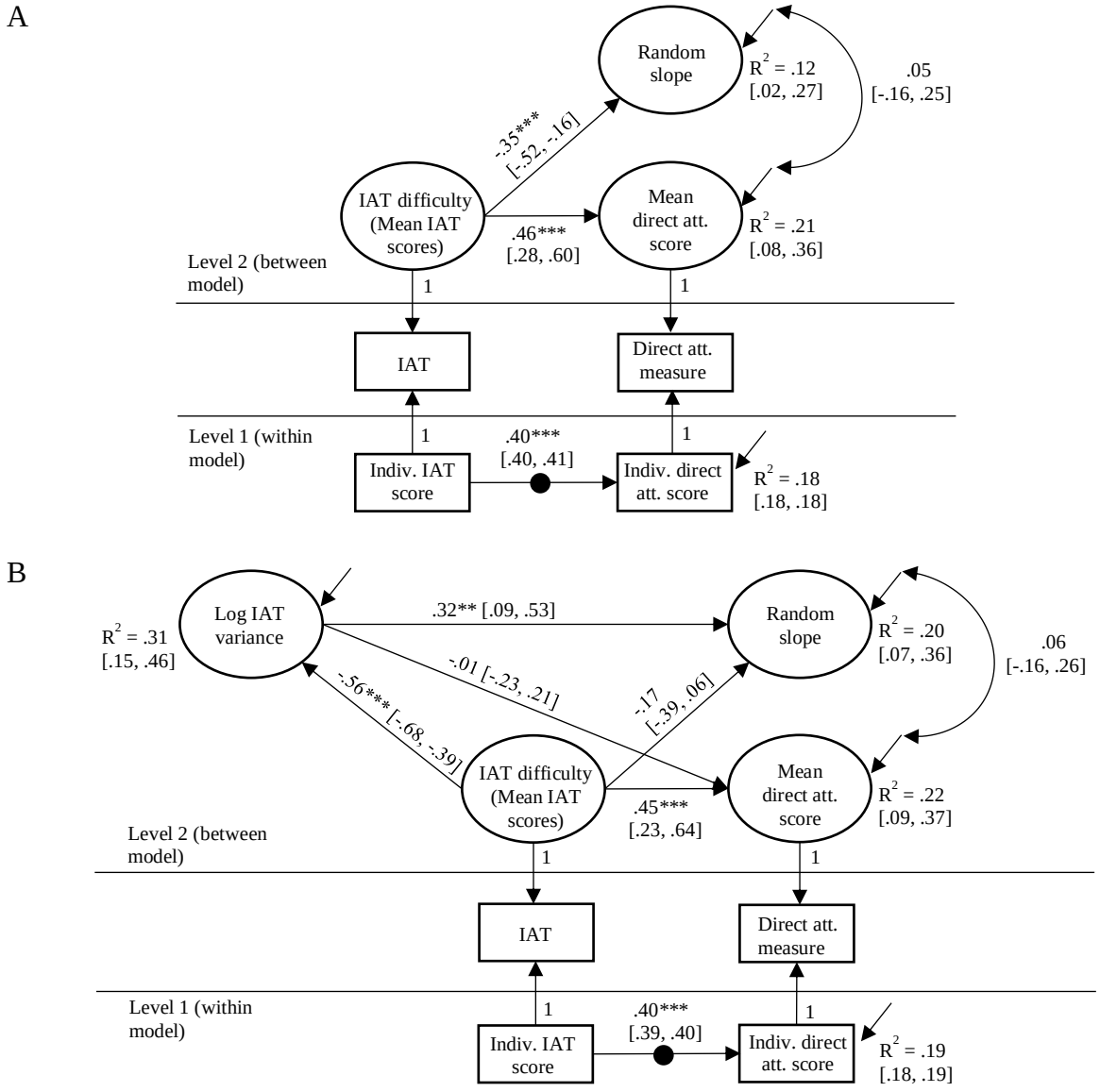
Mediated moderation hypothesis (H2). Standardized results of the full multilevel mediated moderation model are displayed in Figure 2 Panel B (see Supplement 3 for the unstandardized results). The standardized within effect of IAT scores on direct attitude scores averaged across all domains was almost identical to previous results, b = .40, pSD = .002, 95% *C.I.* [.39, .40]. Overall, IAT scores explained 18.5% of variance in direct attitude scores, *C.I.* [18.1%, 18.8%]. The inclusion of log IAT true-score variance as a mediator into the model led to the expected results. The between effect of IAT test difficulty on the random slope, i.e., the cross level interaction of IAT test difficulty and IAT scores was not significant, b = -.17, pSD = .12, 95% *C.I.* [-.39, .06], but the between effect of log IAT true-score variance on the random slope, i.e., the cross level interaction of log IAT true-score variance and IAT scores was, b = .32, pSD = .11, 95% *C.I.* [.09, .53]. Furthermore, the indirect effect of IAT test difficulty on the random slope via log IAT true-score variance was significant as well, b = -.18.[9] Both results suggest that the moderating effect of IAT test difficulty on the relationship

[9] Note that Mplus only provides posterior standard deviations and 95% credible intervals for unstandardized indirect effects.

between IAT scores and direct attitude scores was mediated by log IAT true-score variance. That is, IAT scores have a positive within effect on direct attitude scores that is stronger for IATs of medium difficulty due to an increase in true-score variance and weaker for IATs of extreme difficulties due to a decrease in true-score variance. Figure 4 illustrates this relationship (a list of the estimates by domain can be found in Supplement 3). Together IAT test difficulty and log IAT true-score variance explained 20.1% of variation in the slopes, *C.I.* [7.3%, 35.9%]. Due to the already outlined reasons, we additionally used a meta-analytical approach for better comparisons with other moderators from previous results, according to which both moderators explained 36.9% of variation in correlations between IAT scores and direct attitude scores across domains (for a complete description of the meta-analytical results see Supplement 3).

Figure 2

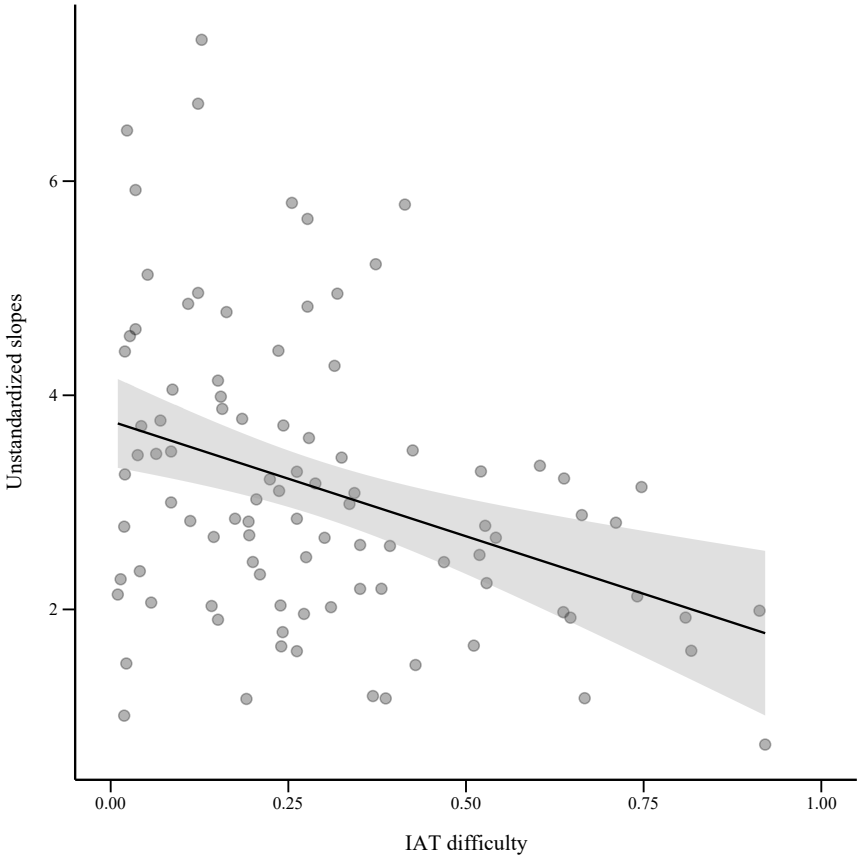
Standardized Parameter Estimates for the Multilevel Moderation Model (Panel A) and the Multilevel Mediated Moderation Model (Panel B) in Study 1



Note. Circles represent latent and rectangles observed variables. Numbers without brackets ascribed to single-headed arrows are path coefficients, numbers without brackets ascribed to double-headed arrows are correlation coefficients, and numbers in square brackets are 95% credible intervals. Darkened circles represent random slopes, which are also depicted on Level 2. R^2 = the coefficient of determination; IAT = implicit association test; Log IAT variance = Log IAT true-score variance; att = attitude; Indiv = individual.
*p < .05, **p < .01, ***p < .001.

Figure 3

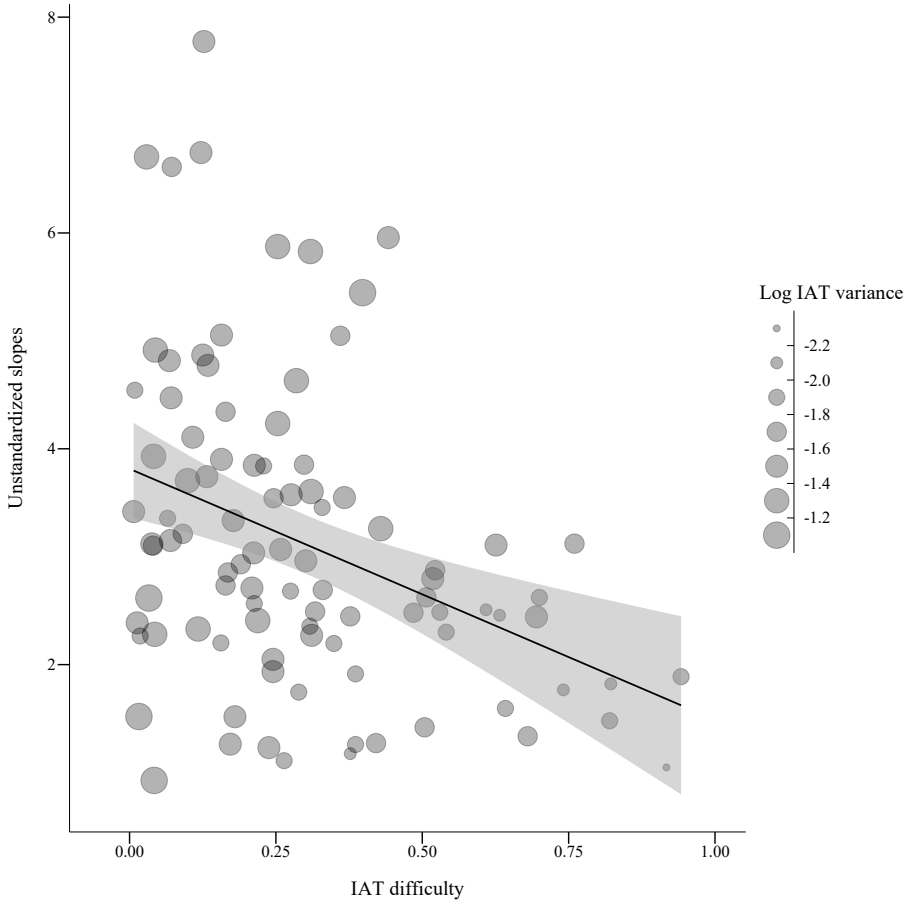
Scatter Plot of the Relationship Between IAT Test Difficulty and Unstandardized Slopes in Study 1



Note. The y-axis (unstandardized slopes) refers to the unstandardized within effect of attitude IAT scores on direct attitude scores for the different domains. The x-axis (IAT difficulty) refers to the latent mean IAT scores of the different domains. The closer the mean IAT scores are to zero the more the IATs approach moderate test difficulty and the further away the mean IAT scores are from zero the more the IATs approach extreme test difficulty. IAT = implicit association test.

Figure 4

Bubble Plot of the Relationship between IAT Test Difficulty, Log IAT True-Score Variance, and Unstandardized Slopes in Study 1



Note. The y-axis (unstandardized slopes) refers to the unstandardized within effect of IAT scores on direct attitude scores for the different domains. The x-axis (IAT difficulty) refers to the latent mean IAT scores of the different domains. The closer the mean IAT scores are to zero the more the IATs approach moderate test difficulty and the further away the mean IAT scores are from zero the more the IATs approach extreme test difficulty. The different bubble sizes (log IAT variance) refer to the log true-score variances of the IAT scores of the different domains. The closer the log true-score variances are to zero the higher the true-score variances. IAT = implicit association test.

Statistical power and other parameter indices of our analyses

We ran a Monte Carlo simulation study for each of the fitted models to validate our results. Both simulation studies showed low parameter estimation bias ($PEB \leq .10$) except for three cases ($-0.33 \leq PEB \leq -0.14$), which was due to small population parameters, low standard error bias ($SEB \leq .11$), small mean squared errors, and high coverage rates close to

.95. All significant effects that have been reported in the previous section revealed a statistical power of .77 or greater. Non-significant effects found in the applications showed lower statistical power in the simulation studies, but mostly still low parameter estimation bias, low standard error bias, and high coverage rates. The results of the simulation studies indicate overall good statistical performance of the models (see Supplement 3 for a complete description of the results of the simulation studies). Accordingly, our efforts to ensure statistical power, firstly by turning to the AIID study, which provided us with a large number of IATs as well as participants per IAT, and secondly by selecting direct attitude measures with few missing values, were successful.

Discussion

In our first study we provided empirical evidence supporting our considerations regarding the application of test difficulty to the IAT by investigating the relationship between attitude IATs of different test difficulties and corresponding direct attitude measures in a subset of the AIID study.

In a first step, we demonstrated that the test difficulty of attitude IATs is a powerful moderator of the relationship between attitude IATs and direct attitude measures, and thus strongly influences the predictive power of attitude IATs. More specifically, the moderation analysis showed that the predictive power of an IAT increases the more the IAT approaches moderate difficulty, that is, the closer the average IAT effect is to zero, and that the predictive power of an IAT decreases the more the IAT approaches extreme difficulty, that is, the further away the average IAT effect is from zero. This was also true when we controlled for complementarity and social sensitivity, two variables describing the domains that have been discussed as relevant moderators of the predictive power of IATs in previous research (Greenwald et al., 2009; Kurdi et al., 2019). In a second step, we demonstrated that this moderating effect of test difficulty was in turn mediated by the true-score variance in IAT effects of attitude IATs. More specifically, the mediated moderation analysis showed that the

predictive power of an IAT increases the closer the IAT is to medium difficulty due to an increase in true-score variance and that it decreases the closer the IAT is to extreme high or low difficulty due to a decrease in true-score variance. In other words, we demonstrated that IAT test difficulty affects IAT true-score variance and thus the predictive power of IATs. As such, we have provided initial evidence for the validity of the test difficulty concept for the predictive power of attitude IATs, and at the same time demonstrated that our application of test difficulty to the IAT is not only reasonable from a theoretical, but also from an empirical standpoint, as the results are consistent with what CTT predicts.

So far, however, we have tested our ideas only in the context of attitude IATs. Therefore, in order to extrapolate our ideas to IATs that measure constructs other than attitudes, we conducted a second study where we tested our ideas in the context of identity IATs. In doing so, we not only test the generalizability of our previous findings to other to-be-measured constructs, but also whether our results can be replicated with other samples and in other contexts.

Study 2

The aims of Study 2 were a) to test the generalizability of the results from Study 1 and b) to replicate the results from Study 1. Therefore, we again investigated whether IAT test difficulty moderates the relationship between IATs and outcome variables (H1) and whether this effect is mediated by IAT true-score variance (H2) in the proposed ways, but this time we used different to-be-measured constructs, samples, and contexts. To this end, we resorted to a different subset of the AIID Study that consisted of the same 95 domains, but solely of identity IATs instead of attitude IATs.

Method

Design and procedure

As our second study also drew on the AIID study, the design and procedure was again consistent with that of the AIID study (see the design and procedure section of Study 1 for a more detailed description).

Sample

As in Study 1, we differentiated between IATs as observations and participants as observations. With respect to the IATs, the study consisted of 95 identity IATs, one IAT for each of the 95 domains. With respect to the participants, the study consisted of 46,045 observations in total (exploratory as well as confirmatory dataset of the AIID study). After excluding participants to ensure data quality (we used the same criteria as described in the results section of Study 1), the overall sample consisted of 43,745 participants. The number of participants was relatively evenly distributed among the different IATs ranging from 303 to 575 with an average of 460 participants per IAT (*SD* = 50.77). In terms of demographic criteria the sample was very similar to our first study and thus rather diverse overall: from those who gave information on the corresponding demographic data 57.9% were 30 or younger, 40% were between 31 and 60, and 2.1% were over 60; 65.3% were female and 34.7% were male; 77.6% were from the United States and of the 22.4% non-United States respondents about half came from Australia, Britain or Canada; 53.9% had a university degree, 34.8% had a college or associate's degree, and 11.3% reported having a high school diploma or less education.

Measures

Indirect Measures: IATs. The identity IATs were similar in procedure and structure to the attitude IATs in Study 1, with the following differences: identity IATs contained a) the attribute categories self/other and corresponding self/other attribute stimuli instead of valenced attribute categories and stimuli, b) five instead of six attribute stimuli, and c) a single set instead of seven sets of attribute stimuli, which was the same for each domain. For

more information on the single identity IATs, see the OSF page of the AIID study again (Link: https://osf.io/pcjwf/).

The IAT effects were calculated in the same way as the IAT effects of the attitude IATs in Study 1 (see the measures section in Study 1 for a detailed description).

Direct Measures: Gut Reactions, Actual Feelings, and Preferences. We used exactly the same self-reported attitude items as in study 1 and also calculated the final outcome variable in the exact same way (see the measures section in Study 1 for a detailed description).

Data Analysis

Since the study design is the same as in Study 1, the data structure is also the same, i.e., a multilevel data structure in which the individual scores are nested within the 95 different domains. It further follows from the same study design that participants were allowed to take part repeatedly, as described in Study 1. In principle, this would once more allow the use of cross-classified multilevel models, but we decided against this possibility again. Instead, we opted for traditional multilevel models, in order to establish comparability with the previous analyses from Study 1. Thus, we used the same multilevel models to replicate the results of both our moderation (H1) and our mediated moderation (H2) hypotheses (see the data analysis section of Study 1 for a detailed description of the multilevel models).

Transparency and Openness

Since study 2 was also based on the AIID study, research ethics committee approval was granted, and all study material can again be viewed on the AIID's OSF page (Link: https://osf.io/pcjwf/). The exploratory and confirmatory dataset were once again the starting point for all further analyses specific to Study 2. The same statistical software as in Study 1 was used for the respective analyses. All corresponding code can be found on our OSF page (Link: https://osf.io/ex9ar/). Since we used the same criteria for data exclusion as in Study 1, we refer the reader to the results section of Study 1 for more information. Information on

statistical power can be found in the results section of Study 2. As Study 2 is a replication of Study 1, the hypotheses and analysis plans were identical. Accordingly, the same preregistration applies, which can be found on our OSF page.

Results

Preliminary analyses

Ensuring data quality. We used exactly the same criteria to ensure data quality as in Study 1 (see the corresponding section of Study 1 for a detailed description).

Assessing the need for multilevel models. Using the same statistics as in Study 1, we again found that multilevel modeling was needed (see Supplement 4 for results).

Bayesian analysis, handling missing values, and descriptive statistics. We used the same estimation methods as in Study 1 (see the corresponding section of Study 1 for a detailed description). In the present study the independent variable individual identity IAT scores had no missing values due to the criteria applied to ensure data quality, but the dependent variable individual direct attitude scores had 6.63% missing values.[10] Descriptive statistics of the observed variables can be found in Supplement 4.

Main analyses

Moderation hypothesis (H1). Standardized results of the full multilevel moderation model are displayed in Figure 5 Panel A (unstandardized results can be found in Supplement 4). The standardized within effect of identity IAT scores on direct attitude scores averaged across all domains was significant, with b = .38, a posterior standard deviation of pSD = .004, and a 95% credible interval of *C.I.* [.38, .39]. Overall, identity IAT scores explained 17% of the variance in direct attitude scores, *C.I.* [16.4%, 17.6%]. In other words, the higher the identity IAT scores the higher the direct attitude scores of the participants on average. As expected, the between effect of IAT test difficulty on the random slope, i.e., the cross level interaction of IAT

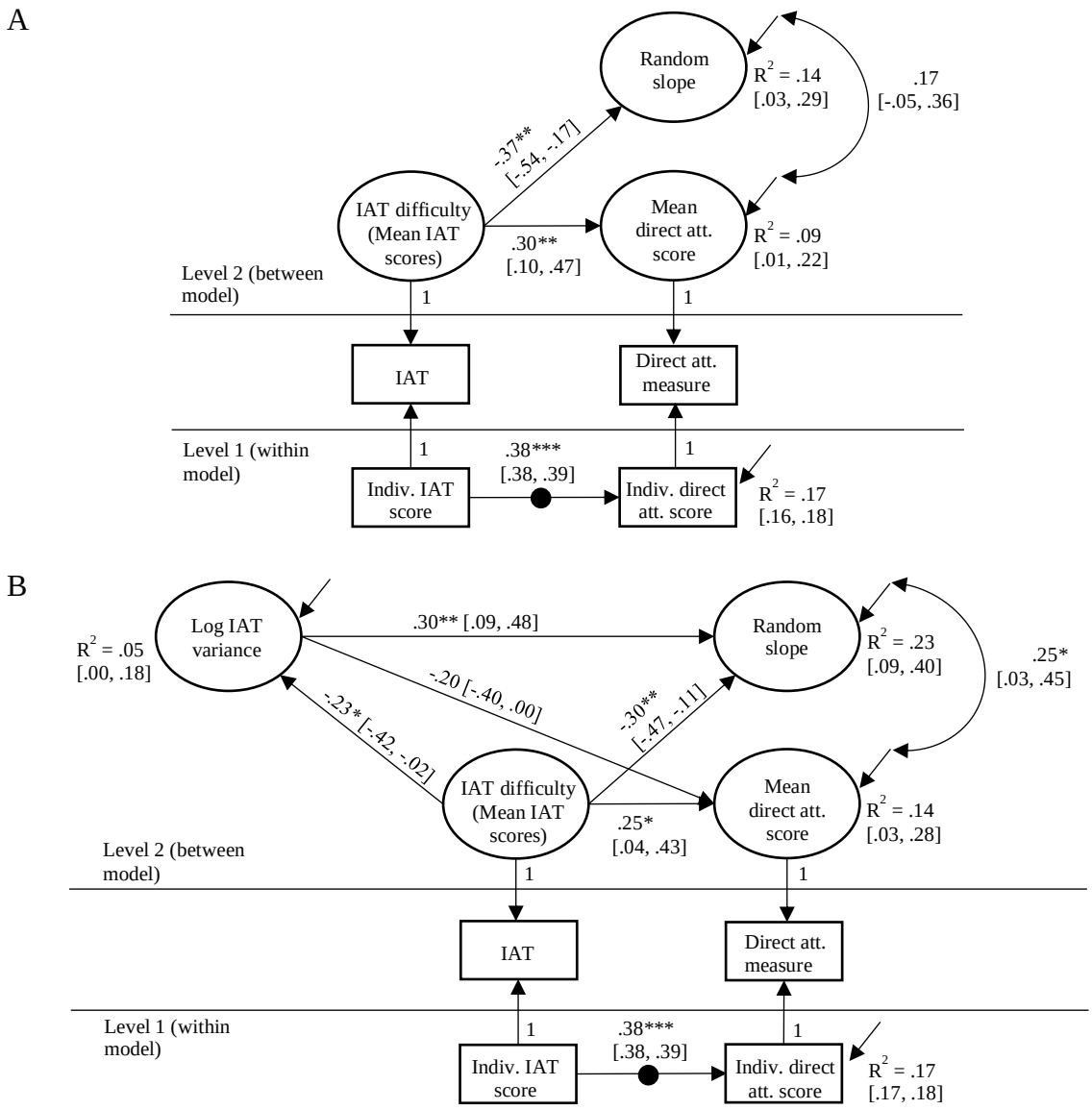
[10] Note that cases were given missing values on the final outcome variable as described in Study 1.

test difficulty and identity IAT scores, was also significant, with a standardized regression coefficient of b = -.37, pSD = .09, 95% *C.I.* [-.54, -.17], suggesting a moderating effect of IAT test difficulty. Thus, in line with our hypothesis H1, the positive within effect of identity IAT scores on direct attitude scores is stronger for IATs the closer their mean D score is to zero (IATs of medium difficulty) and weaker for IATs the further away their mean D score is from zero (IATs of extreme difficulty). See Figure 6 for an illustration of this relationship (a list of the estimates by domain can be found in Supplement 4).[11] IAT test difficulty explained 13.6% of variation in the slopes, *C.I.* [2.9%, 28.8%], using the Mplus R^2 option. As outlined in Study 1, we additionally used a meta-analytic approach to better compare the explained amount of variance R^2 with that of other moderators from previous meta-analytic results. In this meta-analytic framework IAT test difficulty explained 14.89% of the variance in correlations between identity IAT scores and direct attitude scores across domains (a complete description of the meta-analytical results can be found in Supplement 4).

[11] Note that the figure is once again based on unstandardized estimates since Mplus does not provide the corresponding standardized estimates.

Figure 5

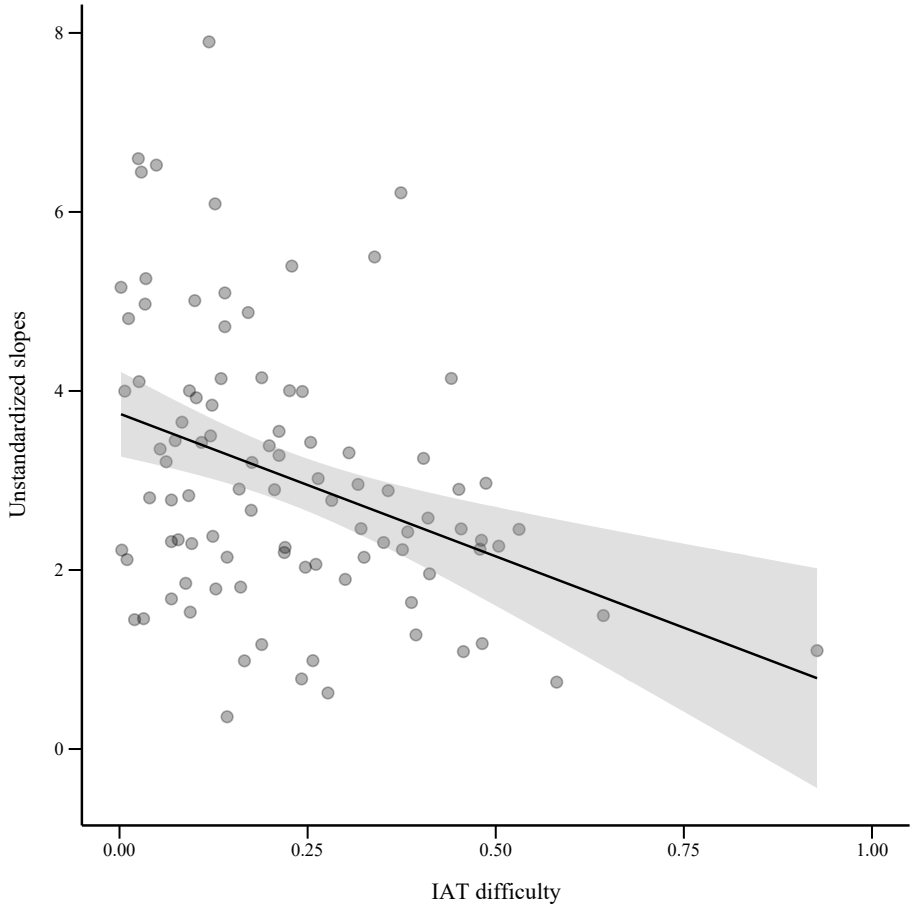
Standardized Parameter Estimates for the Multilevel Moderation Model (Panel A) and the Multilevel Mediated Moderation Model (Panel B) in Study 2



Note. Circles represent latent and rectangles observed variables. Numbers without brackets ascribed to single-headed arrows are path coefficients, numbers without brackets ascribed to double-headed arrows are correlation coefficients, and numbers in square brackets are 95% credible intervals. Darkened circles represent random slopes, which are also depicted on Level 2. R^2 = the coefficient of determination; IAT = Implicit Association Test; Log IAT variance = Log IAT true-score variance; att = attitude; Indiv = individual.
*p < .05, **p < .01, ***p < .001.

Figure 6

Scatter Plot of the Relationship Between IAT Test Difficulty and Unstandardized Slopes in Study 2



Note. The y-axis (unstandardized slopes) refers to the unstandardized within effect of identity IAT scores on direct attitude scores for the different domains. The x-axis (IAT difficulty) refers to the latent mean IAT scores of the different domains. The closer the mean IAT scores are to zero the more the IATs approach moderate test difficulty and the further away the mean IAT scores are from zero the more the IATs approach extreme test difficulty. IAT = Implicit Association Test.

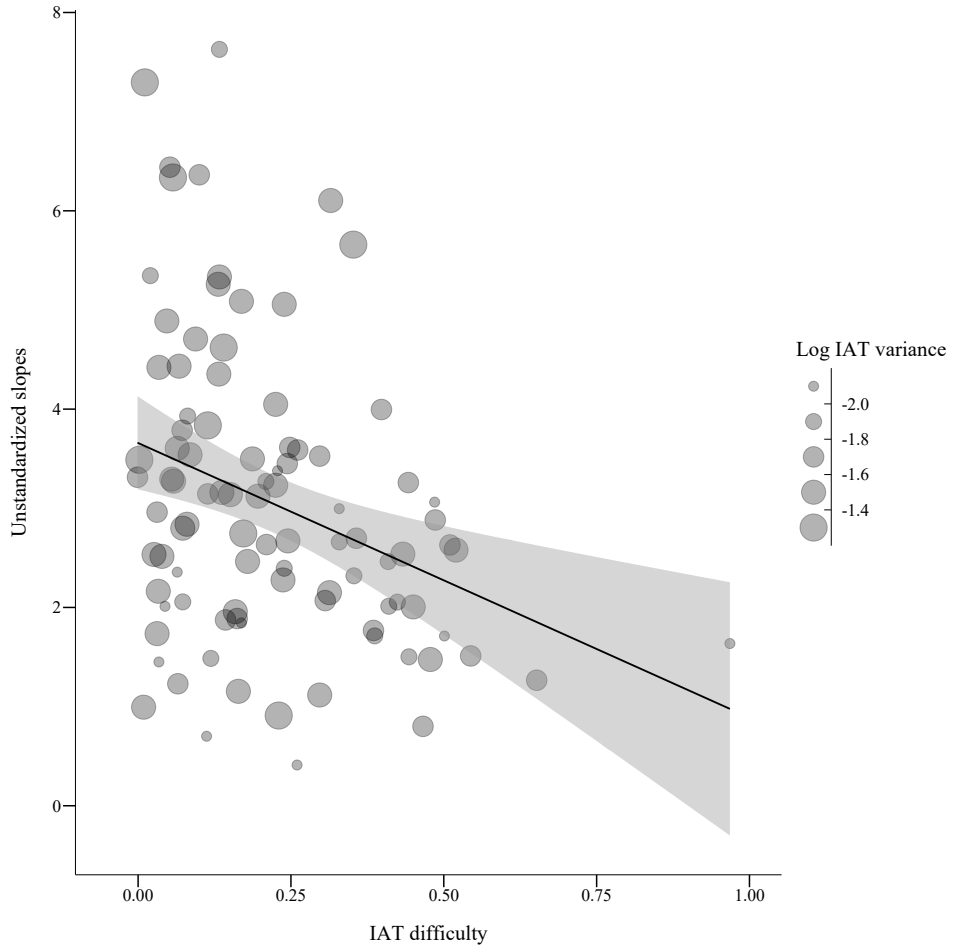
Mediated moderation hypothesis (H2). Standardized results of the full multilevel mediated moderation model are displayed in Figure 5 Panel B (the unstandardized results can be found in Supplement 4). The standardized within effect of identity IAT scores on direct attitude scores, averaged across all domains, was almost identical to that of the moderation model, b = .38, pSD = .004, 95% *C.I.* [.38, .39]. Overall, identity IAT scores explained 17.4% of variance in direct attitude scores, *C.I.* [16.8%, 18%]. The inclusion of log IAT true-score variance as a mediator into the model yielded results predominantly as expected. The between

effect of log IAT true-score variance on the random slope, i.e., the cross level interaction of log IAT true-score variance and identity IAT scores, was significant, b = .30, pSD = .10, 95% *C.I.* [.09, .48], while the between effect of IAT test difficulty on the random slope, i.e., the cross level interaction of IAT test difficulty and identity IAT scores, was reduced but remained significant as well, b = -.30, pSD = .09, 95% *C.I.* [-.47, -.11]. Nevertheless, the indirect effect of IAT test difficulty on the random slope via log IAT true-score variance was significant, b = -.07.[12] The results suggest that log IAT true-score variance mediates the moderating effect of IAT test difficulty on the relationship between identity IAT scores and direct attitude scores. That is, identity IAT scores have a positive within effect on direct attitude scores that is stronger for moderate difficult IATs, at least partly due to an increase in true-score variance, and weaker for extreme difficult IATs, at least partly due to a decrease in true-score variance. See Figure 7 for an illustration of this relationship (a list of the estimates by domain can be found in Supplement 4). Together IAT test difficulty and log IAT true-score variance explained 22.9% of variation in the slopes, *C.I.* [8.9%, 39.5%]. According to the meta-analytical results both moderators explained 39.67% of variation in correlations between identity IAT scores and direct attitude scores across domains (a complete description of the meta-analytical results can be found in Supplement 4).

[12] Note that Mplus only provides posterior standard deviations and 95% credible intervals for unstandardized indirect effects.

Figure 7

Bubble Plot of the Relationship Between IAT Test Difficulty, Log IAT True-Score Variance, and Unstandardized Slopes in Study 2



Note. The y-axis (unstandardized slopes) refers to the unstandardized within effect of identity IAT scores on direct attitude scores for the different domains. The x-axis (IAT difficulty) refers to the latent mean IAT scores of the different domains. The closer the mean IAT scores are to zero the more the IATs approach moderate test difficulty and the further away the mean IAT scores are from zero the more the IATs approach extreme test difficulty. The different bubble sizes (log IAT variance) refer to the log true-score variances of the IAT scores of the different domains. The closer the log true-score variances are to zero the higher the true-score variances. IAT = Implicit Association Test.

Statistical power and other parameter indices of our analyses

We ran a Monte Carlo simulation study for each of the fitted models to validate our results. Both simulation studies showed low parameter estimation bias ($PEB \leq .05$), low standard error bias ($SEB \leq .10$), high coverage rates close to .95, and small mean squared errors. All significant effects that have been reported in the previous section revealed a

statistical power of .84 or greater in case of the moderation model and of .44 or greater in case of the mediated moderation model. The lower statistical power in the latter model was mainly due to the generally lower path coefficients that accompanied the modelled mediation. Nevertheless, the effects still showed low parameter estimation bias, low standard error bias, and high coverage rates. The same was true for the non-significant results with lower statistical power. The results of the simulation studies indicate overall good statistical performance of the models (see Supplement 4 for a complete description of the results of the simulation studies).

Discussion

In our second study, we provided empirical evidence that our considerations regarding the application of test difficulty to the IAT not only hold for attitude IATs but also for identity IATs, again drawing on the large data set provided by the AIID study. We demonstrated in a first step that IAT test difficulty moderates the relationship between identity IATs and direct attitude measures in such a way that the relationship increases the more the IAT approaches moderate difficulty, that is, the closer the average IAT effect is to zero, and that the relationship decreases the more the IAT approaches extreme difficulty, that is, the further away the average IAT effect is from zero. In a second step, we showed that this effect is mediated by the true-score variance in IAT effects in such a way that the relationship between identity IATs and direct attitude measures increases the closer the IAT is to medium difficulty due to an increase in IAT true-score variance and that the relationship decreases the closer the IAT is to extreme high or low difficulty due to a decrease in IAT true-score variance. The fact that our considerations apply not only to attitude IATs but also to identity IATs shows that the difficulty concept can be applied to IATs in general and also replicates our previous findings with different constructs, samples, and contexts.[13]

[13] Note that in Studies 1 and 2, we modeled IAT test difficulty as a latent variable by means of latent cluster means in the multilevel models, but also as a manifest variable in the meta-analytical models and obtained

In the first two studies, we tested our hypotheses drawing on data from the AIID Study, which comprises a large number of different attitude (Study 1) and identity (Study 2) IATs of varying difficulty. The findings of both studies supported our predictions in that IATs of more extreme difficulty had smaller predictive power than IATs of moderate difficulty, and that these effects of test difficulty were due to differences in the variability of the IAT scores. These findings were gathered with a non-experimental approach, since differences in IAT difficulty were merely observed for a given set of different IATs rather than manipulated. Such a non-experimental approach is open to alternative explanations that are due to a potential confounding of IAT test difficulty with other variables that might underlie the effect of test difficulty on predictive power. Although we already controlled for important alternative variables (domain-specific true-score variance in attitudes, social sensitivity, and complementarity), which did not reduce or explain the effects of test difficulty, we cannot completely rule out the possibility that some other variable might be responsible for the effect of IAT test difficulty. To address this concern, we conducted another study (Study 3) in which we experimentally manipulated test difficulty for a given IAT by choosing different reference categories, and then tested whether this resulted in IATs of different difficulty, true-score variance, and predictive power. It is important to note that experimentally manipulating IAT test difficulty is not only important for ruling out potential alternative explanations of our previous findings, but also for researchers who want to design IATs of more predictive power.

Study 3

The aim of Study 3 was to manipulate the test difficulty of a given IAT by choosing reference categories that differ in valence. Given that the IAT is a relative measure of evaluations, in which the evaluation of the relevant target category is compared to the evaluation of a reference category, an IAT should become more difficult the more positive the

similar patterns of results. Consequently, test difficulty can be operationalized as a latent or manifest variable, the latter being consistent with the classic average IAT effect typically reported in IAT research.

valence of the reference category is, and it should become easier the more negative the valence of the reference category is, with IATs approaching moderate difficulty the more similar the reference and the relevant target category are in valence. One approach to modify the valence of the reference categories is to choose reference categories that are independent of the examined study content or context, as was done in the simple association test (SAT) developed by Blanton et al. (2006). In this case, care must be taken that this content-independent reference category is evaluated similarly by all participants in order to not introduce error variance.

For the present study, we chose "environmental protection" as the relevant target category and good/bad as the two attribute categories. Three different IATs were created for this setup by combining the relevant target category with three different reference categories that varied in valence: "environmental degradation" (content-dependent), "war" (content-independent), and "leisure time" (content-independent). The first two reference categories have a strong and unambiguous negative evaluation in comparison to the positively evaluated relevant target category, which should result in very easy IATs (i.e., IATs with large positive average IAT effects) regardless of whether the reference category was content-dependent or not. According to our reasoning, we further hypothesize these IATs to be restricted in their true-score variance and to have little predictive power. The third reference category, "leisure time", has a clear positive valence which is rather similar to the valence of the relevant target category. In this case, participants will be torn between the two target categories. Some will evaluate environmental protection more positively than leisure time or vice versa, while still others will evaluate both similarly positively. This situation should result in an IAT of moderate difficulty (i.e., with an average IAT effect that is close to zero), which should increase the true-score variance attributed to attitudes toward environmental protection, and also its predictive power.

The preceding assumptions can be summarized in the following three predictions: a) The difficulty of an IAT can be influenced by modifying the valence of the reference category. IATs should become more difficult (easier) the more positive (negative) the valence of the reference category, and should approach moderate difficulty the more similar the reference and the relevant target category are in valence. Accordingly, we hypothesize that an Environmental protection/Environmental degradation IAT (Degradation IAT) and an Environmental protection/War IAT (War IAT) are easy and that an Environmental protection/Leisure time IAT (Leisure IAT) is closer to moderate difficulty. b) The manipulation of IAT test difficulty is accompanied by differences in IAT true-score variance. An IAT of moderate difficulty should be better able to discriminate between people on the relevant attitude construct, and should capture more variance attributable to differences in the construct of interest. Accordingly, we hypothesize the Degradation IAT as well as the War IAT to have similar but less true-score variance than the Leisure IAT. c) The manipulation of IAT test difficulty affects the predictive power of the resulting IATs. IATs of moderate difficulty should have more predictive power than IATs of extreme difficulty. Accordingly, we hypothesize the Degradation IAT as well as the War IAT to have similar but less predictive power than the Leisure IAT.

The study was conducted as an online experiment using the recruitment platform Prolific.

Method

Design and Procedure

In a first step, participants were asked to fill out questions about their demographic characteristics. They were then randomly assigned to one out of three different IATs (environmental protection/environmental degradation, environmental protection/war and environmental protection/leisure time). Accordingly we used a between design with the factor *reference category* that had three levels (same content negative valence *environmental*

degradation, different content negative valence *war*, and different content positive valence *leisure time*). The order of the blocks, that is, whether the block in which the relevant target category *environmental protection* and the positive attribute category were on the same response key was the initial combined block or the reversed combined block, was counterbalanced. Upon completion of the respective IAT, participants were asked to answer a series of questionnaire items and exited the experiment.

Sampling and observations

In contrast to Studies 1 and 2 that drew on the AIID study, this time we were not interested in IATs as observations, but only in participants as observations, because the experimental design required other forms of analyses (see Data Analysis section). A total of 481 participants were recruited via Prolific. The requirement for participation was that the participants' native language was German and that they were between 18 and 45 years old. As compensation, the participants received money. Due to exclusion criteria to ensure data quality (see Results section), one participant had to be discarded, resulting in a final sample of 480 participants evenly distributed across the three IATs, with 160 participants per IAT. We aimed for this number of participants based on Schönbrodt and Perugini's (2013) recommendations for a minimum number of participants for stable correlations and based on a one-way a priori power analysis for z-tests of two independent correlations with G*Power (alpha = .05, power = .8, p1 = .2, p2 = .45).

The sample was rather diverse in terms of demographic criteria: from those who gave information on the corresponding demographic data 45.83% were 27 or younger, 37.71% were between 28 and 36, and 16.46% were between 37 and 45 years old; 46.33% were female, 52.20% were male, and 1.47% were diverse; 46.93% were employed, 38.27 received educational training, 7.19% were self-employed, 5.07 were unemployed, and 2.54% were in other occupations; 6.31% studied psychology, 73.65% studied a different subject, and 20.05% did not study/had no university degree.

Measures

Indirect measures: IATs. The following description of the IATs applies to all three IATs. In accordance with the AIID study, we used the standard IAT block structure by Greenwald et al. (2003). The only differences were that the attribute discrimination practice block preceded the target discrimination practice block and that the number of trials for some of the blocks, i.e., all practice blocks, consisted only of half of the trials. The attribute categories were good/bad. The number of attribute stimuli per attribute category was five, and they were the same positive (e.g., love) or negative (e.g., poverty) nouns for all three IATs. The number of target stimuli per target category was also five, but they were always pictures instead of words (e.g., a picture depicting solar panels and wind turbines/car fumes/bombed-out houses/coffee and cake in case of environmental protection/environmental degradation/war/leisure time). All stimuli were presented on a white background. Each stimulus was presented once or twice during each block, depending on the number of trials per block. Attribute and/or target stimuli were presented in random order in each block. Participants were instructed to response as fast and accurately as possible. They had to respond to the stimuli with the left (*D*) or right (*L*) response key. If they had not responded three seconds after the stimulus was presented or did not categorize the stimulus correctly, they received a corresponding feedback that was displayed in red font at the bottom of the page and were asked to continue with the correct response key. All study materials can be found on the OSF page of our project (Link: https://osf.io/ex9ar/).

We calculated IAT effects based on the D score algorithm (Greenwald et al., 2003). The block in which the relevant target category and the positive attribute category share the same response key was defined as the block that constitutes the subtrahend of the difference that produces the IAT effect; positive mean D scores thus indicate a preference of the relevant target category environmental protection over the respective reference category.

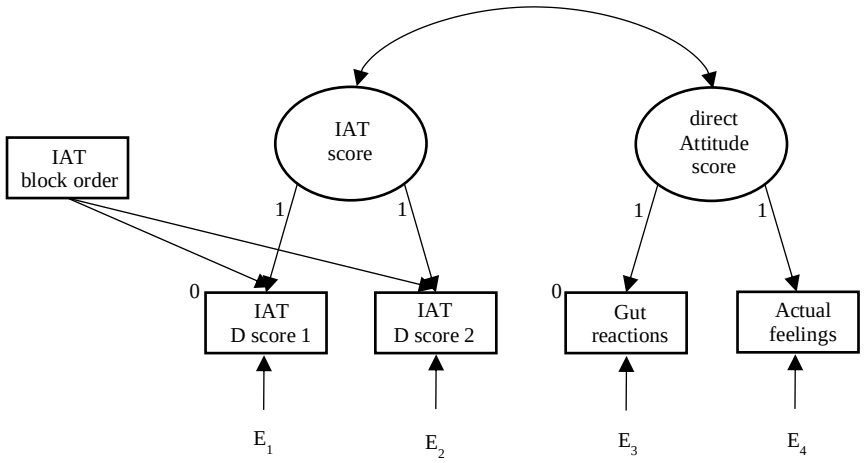
Direct measures: gut reactions and actual feelings. As in Study 1 and 2 we measured gut reactions and actual feelings towards the target categories via self-reported items. For the present study, only the items measuring gut reactions and actual feelings towards the relevant target category environmental protection were used to create an identical outcome variable predicted by each IAT, so that the three different IATs could be compared. Items measuring gut reactions and actual feelings towards the reference categories were not included into the outcome variable, and thus do not contribute to differences in the predictive power of the IATs. Accordingly, the items under consideration were: a) "Rate your gut reactions towards environmental protection, and b) "Rate your actual feelings towards environmental protection". Both items were to be rated on a 10-point bipolar scale with endpoints ranging from 1 *(strongly negative)* to 10 *(strongly positive)*. Higher scores on the items indicate a more positive evaluation of environmental protection.

Data analysis

To test whether modifying the reference categories affected test difficulty, true-score variance, and predictive power of the resulting IATs in the predicted way, we applied multigroup structural equation modeling (SEM; see Breitsohl, 2019; Ployhart & Oswald, 2004), where the manipulated IATs served as three experimental groups. Figure 8 displays the correlated two-factor model that was fitted in all groups. We fitted the model using the R package lavaan (Rosseel, 2012) and the default parameterization of SEM (i.e., the first loading per factor was set to 1 and the first intercept per factor was fixed to zero). The latent IAT factor was measured by two manifest variables: the first IAT D score calculated via the short test blocks and the second IAT D score calculated via the long test blocks. The latent direct attitude factor was also measured by two manifest variables: the gut reactions and the actual feelings towards environmental protection. Additionally, we controlled for potential block order effects by regressing the manifest IAT variables on the manifest covariate IAT block order. The manifest covariate was centered at the grand mean before the analysis.

Figure 8

Basic correlated two-factor model that was fitted in all groups



Note. Circles represent latent and rectangles observed variables. IAT = implicit association test.

First and foremost, this multigroup SEM approach allowed us to test all our assumptions related to our manipulation hypotheses in a single integrated statistical framework. More specifically, we were able to test group differences between the three IATs in a) latent means to determine whether modifying the reference category led to the expected IAT test difficulties, b) latent variances to determine whether the differences in IAT test difficulties were accompanied by the expected IAT true-score variances, and c) latent correlations to determine whether differences in IAT test difficulties and IAT true-score variances led to the expected correlations.

Transparency and Openness

The study received research ethics committee approval from the University of Jena. All material of the study including the preregistration, stimulus material, raw data, curated data, data cleaning code, and the code for our main analyses is publicly available on the OSF page of our project (Link: https://osf.io/ex9ar/). R (version, 4.0.2, R Core Team, 2021) was used for all analyses. For information on statistical power, see the sampling and observations section, and for information on data exclusion, see the results section.

Results

Preliminary analyses

Ensuring data quality. We used the commonly employed criteria of the D score algorithm to ensure data quality, that is, we excluded participants who were faster than 300 ms in 10% or more of the responses in all of the test blocks combined, and we excluded responses with latencies greater than 10,000 ms (Greenwald et al., 2003).

Descriptive statistics, multivariate normal distribution, and handling missing values. Descriptive statistics for the observed variables can be found in Supplement 5. We used the maximum likelihood mean-variance adjusted (MLMV) estimator, since the observed variables were not (multivariate) normally distributed (all *W*s of the Shapiro-Wilk test ≤ .92, all *p*s < .001; Mardia's skewness = 580.93, $p < .001$; Mardia's kurtosis = 24.00, $p < .001$). None of the observed variables had missing values.

Testing measurement invariance. Four types of measurement invariance (MI) are typically distinguished, forming a nested hierarchy with increasing group equality constraints in the order presented: configural MI, weak MI, strong MI, and strict MI (cf. Millsap, 2011; Widaman & Reise, 1997). We tested all four forms of MI except for configural MI because the correlated two factor model shown in Figure 8 required further restrictions (e.g., equal loadings per factor in all groups) to ensure model identification.[14] The results show that strict MI can be assumed, since none of the chi square difference tests were significant (see Supplement 5 for the results of all model tests) and the fit indices of the strict MI model indicate good model fit, S-B $\chi^2_{\text{strict MI}}$ (27) = 40.06, $p = .051$; $\text{RMSEA}_{\text{strict MI}} = 0.06$; $\text{CFI}_{\text{strict MI}} = 0.95$; $\text{SRMR}_{\text{strict MI}} = 0.09$; $\text{AIC}_{\text{strict MI}} = 3761.6$; $\text{BIC}_{\text{strict MI}} = 3874.3$. Based on the results,

[14] Note that at the suggestion of a reviewer, we ran the analyses again with three indicators for the latent IAT factor by creating three parcels at the trial level, hoping that we could then also test the configural MI model. Unfortunately, the configural MI model still did not converge. This was not because of the number of indicators for the latent IAT factor, but because of the number of indicators for the latent direct attitude factor. When we fixed the second indicator of the latent direct attitude factor to 1, the model converged. All other models used to test our hypotheses, yielded similar results to the modeling with two indicators (the results of the modeling with three indicators that we present in Supplement 5 and the corresponding code can be found on our OSF page).

we can assume that the three different IATs measure the same latent construct, and that it is justified to meaningfully test differences with respect to the latent factor variances, means, and their correlations and thus to test our manipulation hypotheses (cf. Widaman & Reise, 1997).

Main analyses

Manipulation hypothesis IAT test difficulty. The latent means of the factor IAT score and their standard errors for the three groups can be found in Table 1. As expected, the latent mean of the leisure IAT, which was slightly negative, $\hat{\mu}_{\text{leisure IAT}}$ = -0.1, was descriptively closer to a D score of zero than the latent mean of the degradation IAT and the war IAT, both of which were strongly positive, $\hat{\mu}_{\text{degradation IAT}}$ = 1.0 and $\hat{\mu}_{\text{war IAT}}$ = 0.81, respectively. To test whether the latent means differed significantly from each other, we introduced another group equality constraint into the strict MI model, namely that the latent mean IAT scores of the three IATs were equal. The resulting model (means model), had a significantly worse model fit than the strict MI model, as indicated by the fit indices in Table 2. It follows, that the latent means differed significantly between the three IATs. Individual Wald tests of each latent mean difference showed that the leisure IAT differed significantly from the degradation IAT, $W(1) = 687.70$, $p < .001$, as well as from the war IAT, $W(1) = 436.36$, $p < .001$, and that the degradation IAT differed significantly from the war IAT, $W(1) = 43.47$, $p < .001$. Finally, we tested whether the latent mean D scores were significantly different from zero, which was true for all three IATs, all zs ≤ -2.42 or ≥ 32.64, all ps < .05. Taken together, the results show that the test difficulties of the IATs generally turned out as predicted, with the environmental degradation IAT and the war IAT being easy IATs and the leisure IAT being the IAT closest to moderate difficulty, although somewhat tending toward a slightly difficult IAT.[15]

[15] Note that while we predicted the latent mean difference between the degradation IAT and the war IAT in our preregistration (see the preregistration also for an explanation), we did not predict the significant deviation from zero of the leisure IAT, but note further that neither result is relevant to our central hypotheses.

Table 1

Latent means, latent true-score variances, latent correlations, R^2, and reliabilities of the IAT factor for the three groups in the strict invariance model

Group	L mean (*SE*)	L variance (*SE*)	L correlation (*CI*)	R^2	Reliability
Degradation IAT	1.0 (.02)	0.02 (.01)	-.02 (-.32, .33)	0.000	0.38
War IAT	0.81 (.025)	0.04 (.01)	.21 (-.06, .47)	0.044	0.52
Leisure IAT	-0.1 (.04)	0.19 (.02)	.35 (.16, .52)	0.123	0.75

Note. L = latent; CI = bootstrap-bias-corrected confidence intervals; Degradation IAT = environmental protection/degradation implicit association test; War IAT = environmental protection/war implicit association test; Leisure IAT = environmental protection/leisure time implicit association test.

Table 2

Model fit of the different models to test the overall manipulation hypotheses

Model	S-B χ^2 (df)	*p*	RMSEA	CFI	SRMR	AIC	BIC	$\Delta\chi^2$	*p*
Strict MI	40.06 (27)	.051	0.06	0.95	0.09	3761.6	3874.3		
Means	194.52 (31)	<.001	0.18	0.41	0.54	4070.4	4166.4	222.50	<.001
Variances	333.43 (35)	<.001	0.23	0.00	0.97	4338.7	4418.0	164.01	<.001
Covariances	340.46 (37)	<.001	0.23	0.00	0.97	4335.6	4406.6	2.22	.33

Note. S-B χ^2 = Satorra-Bentler scaled χ^2; RMSEA = robust root-mean-square error of approximation; CFI = robust comparative fit index; SRMR = robust standardized root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; MI = measurement invariance; Means = strict measurement invariance model plus equal group means; Variances = strict measurement invariance model plus equal group means and variances; Covariances = strict measurement invariance model plus equal group means, variances, and covariances.

Manipulation hypothesis IAT true-score variance. The true-score variances of the factor IAT score and their standard errors for the three groups can be found in Table 1. As hypothesized, the true-score variance of the leisure IAT was descriptively higher, $\widehat{\sigma}_{\text{leisure IAT}}^2$ = 0.19, than the true-score variance of the degradation IAT, $\widehat{\sigma}_{\text{degradation IAT}}^2$ = 0.02, and the war IAT, $\widehat{\sigma}_{\text{war IAT}}^2$ = 0.04. We tested for significant differences in the true-score variances by adding another group equality constraint to the means model, namely equal true-score

variances in all groups. The resulting model (variances model) fitted significantly worse than the means model, as indicated by the fit indices in Table 2. Accordingly, the true-score variances differed significantly between the three IATs. Individual Wald tests of each true-score variance difference, showed that the leisure IAT differed significantly from the degradation IAT, $W(1) = 54.51$, $p < .001$, as well as from the war IAT, $W(1) = 36.50$, $p < .001$, but that the degradation IAT and the war IAT did not differ significantly from each other, $W(1) = 1.98$, $p = .16$. In sum, as predicted, the leisure IAT, which was the IAT closest to moderate difficulty, was also the IAT with the largest amount of true-score variance.

Manipulation hypothesis IAT predictive power. The latent correlations and R^2 for the three groups can be found in Table 1. As hypothesized the latent correlation of the leisure IAT was descriptively larger, $\hat{r}_{\text{leisure IAT}} = .35$, than the latent correlation of the degradation IAT, $\hat{r}_{\text{degradation IAT}} = -.02$, and the war IAT, $\hat{r}_{\text{war IAT}} = .21$. To further explore the differences between the latent correlations, we introduced a final group equality constraint that we added to the variances model, i.e., equal latent covariances in all groups. The resulting model (covariances model) did not fit significantly worse than the variances model, as indicated by the fit indices in Table 2. Accordingly, the latent covariances between the IATs and the direct attitude measures were not significantly different from each other for the three IATs. However, the latent variances of the IATs differed and thus the latent correlations of the three IATs could not be equal in all groups.

To test our specific hypothesis that the leisure IAT has a higher correlation with the outcome than the other two IATs, while the other two IATs have similar correlations at the same time ($\hat{r}_{\text{leisure IAT}} > \hat{r}_{\text{war IAT}} = \hat{r}_{\text{degradation IAT}}$), we conducted Baysian evaluation of informative hypotheses for SEM using the R package lavaan (Rosseel, 2012) in combination with the R package bain (e.g., Gu et al., 2019; Gu et al., 2018). To use the R package bain it was necessary to assume weak MI. The analysis showed that the Bayes factors $BF_{.c}$ and $BF_{.u}$ were both 16.69 (see Van Lissa et al., 2021 for a better understanding of the notation)

indicating that our hypothesis was 16 times more likely than its complement ($BF_{.c}$) or the unconstrained hypothesis, that is, any other ordering of correlations ($BF_{.u}$), and thus the data provide strong evidence for our hypothesis.

Discussion

In our third study, we manipulated IAT test difficulty experimentally, on the one hand to rule out any alternative explanation of its effect on true-score variance and predictive power and on the other hand, to equip researchers with new strategies to influence IAT test difficulty, true-score variance, and predictive power. IAT test difficulty of an IAT for a given to-be-measured construct (i.e., environmental attitudes) was manipulated by choosing the same relevant target category "environmental protection", but three different reference categories that varied in valence (negative reference category – content-dependent: "environmental degradation"; negative reference category – content-independent: "war"; positive reference category – content-independent: "leisure time"). As predicted by the test difficulty account, we found that both IATs with negative reference categories had low difficulty, low true-score variance, low reliability, and low predictive power, regardless of whether or not the reference category was content-dependent on the relevant target category. The IAT with a similar positively valenced reference category as the relevant target category, however, had medium difficulty, more true-score variance, substantial reliability, and significantly increased predictive power. Thus, our study provides strong experimental evidence for the idea that the predictive power of an IAT measuring a particular construct of interest depends on IAT test difficulty, with IATs of moderate difficulty showing the highest predictive power. The study also provides strong evidence that an easy and efficient way to manipulate IAT test difficulty is to choose reference categories of different valence. IAT difficulty and predictive power of the test can be improved by selecting content-independent reference categories that are of the same average valence as the relevant target category. Of course, when selecting such a reference category, it is recommendable to pick a category for

which individual differences in evaluations are low, to avoid introducing error variance that is unrelated to the construct of interest. Our study also shows that choosing a reference category from the same content-dependent domain but with opposite valence, as is often the case in social psychological research, will typically result in extremely easy/difficult IATs that suffer from low true-score variance and low predictive power.

General Discussion

In this article, we started out with the introduction of the test difficulty concept from CTT to IATs in order to increase the predictive power of IATs, and gave a theoretical explanation for how this difficulty effect comes about by elaborating on the dependency between test difficulty and true-score variance. We then provided empirical evidence supporting our test difficulty account by investigating the relationship between attitude IATs and direct attitude measures in Study 1. In a first step, we demonstrated that the relationship between attitude IATs and direct attitude measures was moderated by the test difficulty of an IAT. More specifically, the moderation analysis showed that the predictive power of an IAT increases the more the IAT approaches moderate difficulty, that is, the closer the average IAT effect is to zero, and that the predictive power of an IAT decreases the more the IAT approaches extreme difficulty, that is, the further away the average IAT effect is from zero. In a second step, we demonstrated that this moderating effect of the test difficulty of an IAT was in turn mediated by the true-score variance of an IAT. More specifically, the mediated moderation analysis showed that the predictive power of an IAT increases the closer the IAT is to medium difficulty due to an increase in true-score variance and that it decreases the closer the IAT is to extreme high or low difficulty due to a decrease in true-score variance. In Study 2, we replicated the just described results with other to-be-measured-constructs, samples, and contexts, thereby generalizing the test difficulty account to IATs as it applies to attitude IATs just as well as to identity IATs. As such, we have provided strong initial evidence for the validity of the test difficulty account. In a final step, we provided

experimental evidence for the effects of test difficulty on true-score variance and predictive power in Study 3, by manipulating the test difficulty of an IAT measuring a particular construct of interest by combining the relevant target category with reference categories of different valence. This last study is again of theoretical importance by providing further proof of concept, and also of practical importance, as it demonstrates the usefulness of the test difficulty account in the development of IATs by suggesting design-related modifications of the IAT in order to increase its true-score variance, or specifically in this case, to optimally assesses attitudes towards a specific construct.

In what follows, we provide a synopsis in which we compile and unify existing knowledge on test difficulty and true-score variance regarding the development of IAT studies and address open questions and promising research directions based on this knowledge in order to best advance correlative IAT research in the future. We then, discuss the limitations of our studies, and end with our concluding remarks.

Increasing the predictive power of IATs based on the concept of test difficulty and true-score variance

Test difficulty – an actual safeguard against counterproductive recommendations for correlational IAT research

As already described in the theory section, understanding the relationship between test difficulty and true-score variance can help researchers to identify recommendations and research practices that are counterproductive for increasing IAT true-score variance of IATs and thus for correlative IAT research. One such recommendation prevailing in the literature is to strive for large and robust IAT effects (e.g., Greenwald et al., 1998; Greenwald et al., 2003; Payne et al., 2005). Given the widespread nature of this recommendation in the literature, this may come as a surprise to many readers, but it is all the more important. Striving for large IAT effects implies that researchers aim at designing extremely easy or difficult IATs, which, as we have shown, leads to reduced true-score variance and ultimately to a restriction in

predictive power. Accordingly, the seemingly common practice in IAT research to strive for large IAT effects, although helpful in experimental research and for demonstrating the existence of general biases in social cognition (e.g., stereotypes about or prejudice against minorities), is counterproductive for the purpose of identifying individual differences and for correlational research that aims at predicting individual differences in relevant outcomes. Our results are in line with previous research examining the effects of large effect sizes for correlational research in implicit measures. For example, Hedge et al. (2018) showed that classic cognitive tasks and tests are often problematic when it comes to examining individual differences. They argue that the very reason why cognitive tasks and tests are classical and popular is that they produce robust effects and the reason for that in turn is low variability in the data. The divergent opinions on whether small or large effects are desirable are due to the difference between experimental and correlational research described earlier. When it comes to experimental research, one is interested in mean differences between conditions, and large effects are desirable because they allow valid conclusions and inferences. Since the IAT has its origins in cognitive experimental research, it is not surprising that the general tenor is that large effects are desirable. However, when it comes to correlational research, one is interested in assessing individual differences and in predicting outcome variables, and then, as our research shows, one should not strive for large IAT effects but, on the contrary, for null-effects, that is, IATs that have moderate difficulty and large true-score variance.

Another recommendation that can be counterproductive for correlational IAT research is to develop complementary IATs. Complementarity was one of the strongest moderators of IAT relations in the meta-analysis by Greenwald et al. (2009), which is why researchers could use complementarity as a guideline for developing IATs with more predictive power, or at least not assume a negative impact of complementarity on the predictive power of IATs. However, such a negative impact of complementarity is exactly what our results suggest under certain conditions. While we found complementarity to be a relevant moderator in our

first study, replicating previous results, we found in our third study that the environmental protection/environmental degradation IAT, a complementary IAT, resulted in a very easy IAT with low true-score variance and little predictive power. Although this result in itself is only a single case, it may be indicative of a more general pattern, namely that complementarity, when it coincides with extreme difficulty, has a negative impact on the predictive power of IATs. Such a conflation of complementarity and extreme difficulty may be more common than previous meta-analyses suggest; at least, various hypothetical examples of such cases can be found (e.g., wealth vs. poverty, war vs. peace, Blacks vs. Whites, old vs. young, etc.). Furthermore, it must be taken into account that many of the complementary IATs are close to moderate difficulty and that this is due to the distribution of the to-be-measured construct in the selected sample. For example, in case of complementary IATs with target categories where some people evaluate the one target category very positively (and the other target category very negatively), while other people evaluate the target categories in just the opposite way, such as the target categories evolution/creationism, gun-control/gun-rights, astrology/science in a representative American sample, the result is an IAT trending towards moderate difficulty. For other target categories, or for the same target categories investigated with other samples where all people evaluate the one target category very positively (and the other target category very negatively), however, complementary IATs will have more extreme difficulties, which will decrease their predictive power. Accordingly, the test difficulty account shows that it depends on the conditions whether complementarity will produce IATs of extreme difficulty, reduced variability and low predictive power or not.

Overall, the discussion shows that applying the concept of test difficulty helps IAT researchers to understand previous findings, which shields them from following research practices that are counterproductive for correlational IAT research.

Factors influencing the test difficulty, true-score variance, and predictive power of IATs

Sample. A first factor influencing the predictive power of an IAT is the sample. The more heterogeneous the sample with regard to the to-be-measured construct the larger the predictive power of the IAT everything else being equal. Consider a Black/White attitude IAT. The IAT will have more predictive power when the sample consists of both Blacks and Whites than when it exclusively consists of Blacks or exclusively of Whites. This is why previous research has recommended to recruit heterogeneous samples to directly increase the true-score variance of IAT scores (Greenwald et al., 2022). Effects of the sample are also in line with the test difficulty account, which assumes the sample to not only influence the true-score variance and the predictive power of IATs, but also their test difficulty. In terms of the given example, this means that with a more heterogeneous sample, a given IAT will tend towards moderate difficulty, higher true-score variance and more predictive power. Thus, the test difficulty account supports the same recommendation made in previous research.

To-be-measured construct. To-be-measured constructs can also influence the predictive power of IATs. The larger the variability on the to-be-measured construct in a population the larger the predictive power of the IAT, everything else being equal. It is important to note that little (large) true-score variance in a to-be-measured construct is typically accompanied by extreme (moderate) difficulty of a corresponding IAT. However, the relationship between IAT test difficulty and IAT true-score variance is a probabilistic one and exceptions may occur (e.g., moderately difficult IATs may also reflect to-be-measured constructs with little true-score variance when the target categories are evaluated neutrally by almost everyone). These exceptions, however, do not contradict the general validity of the test difficulty account. In addition, in most cases for a given variability on the to-be-measured construct in a population, IATs of moderate difficulty will still capture more of this true-score variance than IATs of extreme difficulty. Nevertheless, it is important to keep in mind that only when moderate IAT test difficulty reflects an increased amount of true-score variance in the to-be-measured construct will it increase the predictive power of an IAT.

Depending on the research question, researchers either can or cannot choose the to-be-measured construct arbitrarily. For example, when the research question is to compare the evaluation of Blacks vs. Whites, then researchers are bound to the categories Blacks and Whites, and there is no leeway for selecting different target categories. If, for example, on the other hand, the main research question is to investigate the relation between implicit attitudes and behavior, then we would recommend to select target categories where the variability in these attitudes is expected to be large.

Context. Another factor influencing the predictive power of an IAT is the context in which the test is taken. Such contextual factors might, for example, consist of providing attitude relevant information or activating attitudes due to priming procedures given the to-be-measured constructs are attitudes. In comparison to the factors considered so far, the influence of the context on the predictive power appears to be more ambiguous, since the effects depend on a complex interplay of multiple features. For instance, highlighting a specific piece of information about a target may reduce variability in the resulting IAT scores by focusing the knowledge on the primed attribute among the participants who take the test, thus making them more homogenous. Alternatively, specific contexts may also polarize attitudes when they highlight a piece of information that elicits diverse responses from different groups of participants. Finally, the variability of contexts under which the test is taken, when the context is not a controlled factor of the design, may also influence the true-score variance and predictive power of the test (at least when the variability of the context is related to the to-be-measured construct and the predicted outcome variable is assessed in the same context). Empirical research investigating contextual influences on the IAT stems primarily from the literature on the malleability of IAT effects and the change of implicit biases (for overviews, see Blair, 2002; Lai et al., 2013). In this literature, however, the focus was either on the IATs malleability (i.e., differences in average IAT effects) or on its predictive power – contextual modulations of the relations between test difficulty, true-score variance, and predictive power

were not examined. Based on the test difficulty account, we would predict that contexts that homogenize people with regard to their attitudes should lead to IATs of more extreme difficulty, less true-score variance, and lower predictive power, whereas contextual conditions that lead to a polarization should lead to IAT effects of moderate difficulty, more true-score variance, and higher predictive power. These considerations suggest that it might be promising to prime participants with ambiguous information regarding the relevant target category before taking the IAT, that is, with information that highlights conflicting views and beliefs about the target category in different groups of participants, or that activates a certain source of evaluation in only a subgroup of participants. By pushing different subsamples to opposite poles of the evaluative spectrum, the resulting IAT should have a more moderate difficulty, more true-score variance on the to-be-measured construct, and a higher predictive power, because it captures more of the true variance in underlying beliefs and prejudices. As of now, however, these claims and suggestion have to be regarded as speculative due to a lack of evidence.

Design. A final important influencing factor of the predictive power of the IAT is the design of the task itself. If an IAT is constructed in such a way that it results in the IAT tending toward moderate difficulty and higher true-score variance, the predictive power of the IAT increases, everything else being equal. Based on the concept of test difficulty we propose three approaches to modify the design of IATs to influence the test difficulty, true-score variance, and predictive power of IATs: (a) via the reference category, (b) via the attribute categories, and (c) via the exemplar stimuli.

Modifying the reference category. So far we have tested only the first of these approaches empirically, i.e., modifying the design via the reference category (Study 3). Given a researcher is mainly interested in the evaluation of one target category, the design of the IAT can be modified by choosing a reference category of similar valence to the relevant target category which is independent of the examined study content or context and which is

evaluated similarly by the participants of the sample. We have demonstrated the viability of this approach for an environmental attitude IAT as an example in Study 3 above. Whereas the usual construction procedure of the IAT (environmental protection/environmental degradation IAT) led to an easy IAT with low true-score variance, low reliability, and low predictive power, our proposed modification of the reference category based on the test difficulty account (environmental protection/leisure time IAT) led to an IAT with moderate difficulty, with large true-score variance, high reliability, and high predictive power. Of course, the study was just a first example to illustrate the applicability and viability of the test difficulty approach, and further tests with other relevant target categories are necessary to demonstrate the generalizability of this strategy.

Modifying the attribute categories. Another promising approach to manipulate the test difficulty of an IAT regards the choice of the attribute categories, especially in the case of attitude IATs. In the construction of attitude questionnaires, the polarity of evaluative adjective pairs can be changed depending on whether the attitude object under study is a priori thought to be evaluated as extremely positive/negative or not. For attitude objects that are assumed to elicit positive as well as negative evaluations, bipolar adjective pairs such as good/bad are used to cover the expected spectrum of attitudes, whereas when attitude objects are assumed to elicit mostly positive or mostly negative evaluations, unipolar adjective pairs such as good/very good in the case of a positive spectrum of attitudes or bad/very bad in the case of a negative spectrum of attitudes can be used. Choosing adjectives in line with the to-be-expected spectrum of evaluations for an attitude object should make it more likely that the items tend towards moderate difficulty and higher true-score variance, while still adequately covering the attitudes toward the object under study.

Applying this strategy to the IAT to deal with to-be-measured constructs that are a priori thought to lead to IATs with extreme difficulties and low true-score variance, we would assume that unipolar attribute categories (e.g., good vs. excellent) are more likely to result in

IAT effects closer to moderate difficulty, whereas attitude IATs with bipolar attributes categories (e.g., good vs. bad) are more likely to result in IAT effects closer to extreme difficulties. To better understand the basic principle of our idea let us consider a somewhat simplified situation without considering a reference category, in which a single target IAT (Bluemke & Friese, 2008) or a single category IAT (Karpinski & Steinman, 2006) is employed to assess individual differences in attitudes toward a target category that is evaluated positively by a majority of participants (like environmental protection, or health, or financial success). Choosing the attribute categories good and bad will then result in a very easy IAT, since nearly all participants will respond faster in the block in which environmental protection and good share the same key, compared to the block where environmental protection and bad share the same key. Now consider the same IAT except that the attribute categories are exchanged by the categories excellent and good. This modification should result in markedly weaker IAT effects, since a substantial number of participants will now have a tendency to classify environmental protection with good (which now is the more "negative" attribute category) rather than with excellent (the more positive attribute category). If calibrated in a way that the average valence of the target falls in between the two attributes the result should be an IAT of moderate difficulty that captures more of the true-score variance in the to-be-measured attitude construct, and has better predictive power.

To our knowledge, no systematic empirical test of the idea has been reported yet. Kurdi et al. (2019) did examine the polarity of attribute categories as a moderator of the predictive power of IATs and found that higher polarity was associated with higher predictive power. However, high polarity (e.g., fat vs. thin) and low polarity (e.g., sad vs. angry) are not the equivalent to bipolarity vs. unipolarity, nor did they analyze whether the choice of attribute pairs interacted with the positivity/negativity of the relevant target category, which would be the appropriate test of the prediction based on the test difficulty account.

Another issue that comes along with our proposed approach is that it is questionable whether unipolar attribute categories are mutually exclusive (e.g., everything that is excellent could also be considered to be good). This problem can be circumvented by clear instructions (e.g., highlighting that "good" stands for "good but not excellent") and by selecting exemplar stimuli that are unambiguously categorizable to the attribute categories. In summary, modifying the typically bipolar attribute categories toward unipolar attribute categories in the case of attitude IATs that are used to diagnose differences among a spectrum of attitudes that is almost entirely positive (or entirely negative) may be a promising approach to explore in the future.

Modifying the exemplar stimuli. Another possibility to manipulate the test difficulty of an IAT regards the selection of exemplar stimuli for the relevant target and for the reference target category. An IAT can be made more difficult (easier) by choosing more negative (more positive) exemplars for the relevant target category, and/or by choosing more positive (more negative) exemplars for the reference category. These strategies, however, come with a certain risk that participants may (a) redefine the target and reference categories (Govan & Williams, 2004), or that (b) the resulting IAT effects may be more strongly influenced by evaluations for the exemplars than for the categories (e.g., Gast & Rothermund, 2010; Steffens & Plewe, 2001). By selecting representative neutral exemplars or a representative mixture of positive and negative exemplars for both target categories, the effect of the exemplars on the test difficulty should be held constant at a moderate difficulty level, a strategy that would also be consistent with what has already been proposed for the selection of exemplar stimuli (Greenwald et al., 2022) on the basis of preventing recoding (e.g., Gast & Rothermund, 2010) or unwanted representations of the target categories (Govan & Williams, 2004).

Practical implications for the most prominent field of IAT research: Applying the test difficulty concept to IATs assessing implicit prejudice against and stereotypes of social groups

We will now discuss how test difficulty manipulations can be used to increase the true-score variance and in turn the predictive power of IATs that aim at assessing implicit prejudice against and stereotypes of social groups (e.g., immigrants, people of color, old/young, males/females). Basically, we have to distinguish between two different types of research questions in this area: First, IATs can be used to demonstrate *general biases at the population level* by comparing evaluations or attributes between two target groups. In this case, the research question determines the choice of target categories, and IAT test difficulty (i.e., the average D score) is the outcome and reflects the result of the study, leaving no room for design-based manipulations of IAT test difficulty.[16] A second group of research questions aims at assessing *individual differences in attitudes or stereotypes* in order to relate the IAT scores to some outcome variable of interest (e.g., personality traits, other attitudes [explicit or implicit], observable behavior, decision making). In these cases, the true-score variance of the IAT in question should be maximized by designing the IAT in such a way that it has moderate difficulty (i.e., that the average IAT effect is close to zero).

Based on the results of our third study, we would recommend to choose a reference category that has similar valence compared to the relevant target category. For example, if individual differences in implicit prejudice against a specific minority are to be measured (in order to predict discriminatory attitudes or behaviors towards members of this group), then we would recommend to compare the group in question to a reference group that has a similar average evaluation as the relevant target group. In addition to having a similar average

[16] If the aim is to determine absolute evaluations for one (or both) of the target categories, researchers can use the ReAL model (Meissner & Rothermund, 2013) of the IAT to estimate separate parameters indicating these evaluations separately. Again, however, the choice of (reference) target categories is not something that researchers should want to manipulate in this line of research.

evaluation as the relevant target group, attitudes towards the reference group should also be uncorrelated with attitudes towards the relevant target group, and should show only little variance within the population of interest. For example, if the aim is to assess individual differences in implicit attitudes towards, say, gay people, and one expects slightly negative evaluations of this group in the population one is focusing on (say, middle-aged people in a Republican-dominated state), then it might be good to choose a reference category like "older people" or "Hispanics", which is similarly evaluated (slightly negatively) in this population, is unrelated to the attitude in question ("gay people"), and for which there is relatively little variance in attitudes among the participants of the study.

Similarly, when the aim is to assess individual differences in implicit stereotypes of a particular group (i.e., associations between the group and a specific attribute), then we would recommend to choose a reference group in the IAT that is – on average – similarly evaluated on the respective attribute dimension as the relevant target group. For example, if the aim is to assess differences in the implicit stereotype of male intelligence, then "men" (the relevant target category), which may be evaluated as being slightly above average on intelligence, could be compared to a reference category like "Asians", which also has a connotation of being above average on intelligence that is unrelated to the stereotype of men, and also has little variance among the population of interest (e.g., female Europeans).

Limitations

Possible confounders and alternative explanations for our results

With respect to the first two nonexperimental studies, questions arise about possible confounding variables responsible for the relationship between IAT test difficulty and the IAT's predictive power. For this discussion, it is first important to distinguish between the influencing factors of test difficulty that we already discussed, and possible confounding variables. Influencing factors have an effect on test difficulty, which in turn mediates the effect of these factors on the predictive power of IATs. While the former is also true for

confounding variables, the latter is not. Consequently, variables such as characteristics of the context, the sample, or the to-be-measured constructs cannot be considered confounding variables if they are also considered influencing factors that affect the test difficulty of an IAT. However, in addition to theoretical considerations to establish whether the variables are influencing factors or confounding variables, it is important to ensure empirically that the variables are not confounding variables by showing that IAT test difficulty has an additional effect on the predictive power of IATs. Due to the random assignment of the participants to the IATs in the AIID study, it can be assumed that characteristics of the sample or the context are unrelated to the predictive power of IATs, thus ensuring an additional effect of IAT test difficulty over and above these variables. The to-be-measured constructs, on the other hand, are perfectly intertwined with the IATs, as the domains always equal the target categories of the IATs in the AIID Study. Therefore, we took a closer look at the characteristics of the to-be-measured constructs. In our first study, we demonstrated that IAT test difficulty explained the predictive power of IATs over and above the moderators social sensitivity and complementarity, both of which had been identified as possible moderators in previous meta-analyses and describe the to-be-measured constructs (Greenwald et al., 2009; Kurdi & Banaji, 2019; Kurdi et al., 2019). We also examined whether IAT test difficulty can explain the predictive power of IATs over and above the variance of the to-be-measured constructs. Certain domains might, for example, simply be more controversial than other domains and thus inherently have more true-score variance. We used the true-score variance of the direct attitude measures as a proxy for the true-score variance of the to-be-measured constructs and analyses showed that IAT test difficulty did not correlate with the true-score variance of the direct attitude measures and thus had an additional effect on the predictive power of IATs (see Supplement 3 for conceptual representations of the models and the corresponding standardized results). Other possible confounding variables discussed in the literature concern the outcome variable. However, due to the design of the AIID Study and our choice of the

final outcome variable, these moderators are constant across the different domains and their corresponding IATs and therefore cannot explain our findings. These moderators include, but are not limited to, the *correspondence between IATs and outcome variable*, *controllability of responses to the outcome variable*, *subject responses to the outcome variable vs. experimenter-observed outcome variable*, *conscious awareness about the hypotheses* (Greenwald et al., 2009; Kurdi et al., 2019).

In Study 2, we were able to replicate the results of Study 1 with different to-be-measured constructs, that is, identities instead of attitudes, and in doing so, give further evidence that the results of Study 1 do not reflect the operation of hidden confounding variables. Note that participants were also randomly assigned to the IATs in Study 2 and that the final outcome variable was the same as in Study 1, which also reduces the possibility of an influence of confounding variables on the results of Study 2.

Last but not least, we experimentally manipulated IAT test difficulty within a given to-be-measured construct in Study 3, and we were again able to show that changes in test difficulty result in changes in true-score variance and predictive power. In theory, this experimental design eliminates the influence of all confounding variables. However, it does not exclude the effect of variables that are inherently intertwined with the manipulation (or design of the study) and that might mediate or cause the effects of test difficulty, which we argue is the case for recoding. Recoding describes the process of converting the four nominal categories of an IAT into two superordinate categories based on features that allow for such simplification, such as valence in an attitude IAT, in the compatible block, with the result that the exemplar stimuli are not sorted based on their nominal category membership but on the respective superordinate categories (e.g., positive/negative in case of the feature valence; Mierke & Klauer, 2003; Rothermund et al., 2009). With respect to our experimental study, it seems likely that the easy IATs allow for recoding because the target category environmental protection is positive, whereas the target category environmental degradation or war is

negative, so that both target categories share keys with attribute categories of the same valence in the compatible block (positive and negative, respectively). The moderately difficult IAT does not allow for recoding because both target categories (leisure time and environmental protection) are positive. We examined the role of recoding in more detail in Study 3: Based on task-switch-costs analyses we calculated an indicator for recoding (i.e., the difference between the task-switch-costs in the incompatible and the compatible block; Mierke & Klauer, 2001, 2003), which we subsequently included as an additional predictor in our multigroup SEM analyses (the results of these analyses are presented in Supplement 5 and can be viewed on our OSF page along with the corresponding code). The main result was that controlling for recoding eliminates the difference between the latent correlations of the moderate difficult IAT (leisure IAT) and the easy IATs (war and degradation IAT). Evidently, recoding affects the predictive power of IATs: In easy IATs, where recoding is present, construct-irrelevant variance is introduced and IAT results become distorted due to recoding, because recoding eliminates the influence of the target categories that are no longer used for performing the task, and as a result, the predictive power of the IATs is reduced. In moderately difficult IATs, on the other hand, recoding is difficult or hardly feasible, eliminating construct-irrelevant variance attributable to individual differences in recoding and allowing the true evaluation differences of the target categories to influence task performance. It is important to note, however, that the exact causal role of recoding remains unclear. Recoding could be a causal variable affecting test difficulty or it could be a mediator explaining the effects of test difficulty. Based on our study design, we cannot disentangle these possibilities because test difficulty and recoding are inherently intertwined. In order to specify the causal role of recoding and test difficulty, experiments are needed that allow for a separate manipulation of recoding and test difficulty. It is also important to bear in mind that we have so far examined only one experimental study with respect to the role that recoding plays for the predictive power of IATs. Unfortunately, the role of recoding cannot be

examined in Studies 1 and 2 because the target-exemplars and the attribute-exemplars were presented in a strictly alternating sequence in project implicit, making it impossible to compute task-switch-costs (all trials are switch trials). Consequently, to understand the exact relationship between recoding and test difficulty, true-score variance, and predictive power, further studies with conditions different from those in our Study 3 are needed. While we consider this to be an insightful and important endeavor, it is critical to point out that regardless of the exact causal role that recoding ultimately plays, it does not call into question the validity and usefulness of the test difficulty concept; rather, it provides invaluable insights for a more comprehensive understanding of the underlying processes at play.

Relevance of IAT test difficulty for the predictive power of IATs

Although many of the moderators of the IAT's predictive power discussed in the literature, especially moderators that affect the outcome variable, were constant in all our studies, ruling them out as possible confounders, it remains an open question how important IAT test difficulty is relative to these moderators in explaining the predictive power of IATs. To ensure comparisons to previous results, we adopted a meta-analytical approach in addition to our multilevel approach in Studies 1 and 2, since the former is more in line with how moderators of the IAT's predictive power were investigated so far (cf. Greenwald et al., 2009; Kurdi & Banaji, 2019; Kurdi et al., 2019). These meta-analytical results suggest that IAT test difficulty is a relevant moderator explaining between 15% and 24% of the variance in the predictive power of IATs. The importance of IAT test difficulty is also evident when compared to the additional moderators that we could investigate, that is, social sensitivity and complementarity. The analyses in Study 1 showed that IAT test difficulty explained more variance in the predictive power of the IATs than the two moderators combined. Notably, complementarity was one of the strongest moderators in previous research when included in the analyses (Greenwald et al., 2009), and IAT test difficulty outperformed complementarity (it should also be noted here that we have also provided arguments and evidence [Study 3]

questioning the benefit of the moderator complementarity for developing IATs with more predictive power). Finally, the relevance of IAT test difficulty is further confirmed by Study 3, where the moderate difficult IAT explained 12% more of the variance in the direct attitude measures compared to an IAT that would typically be designed without considering the test difficulty account.

Predicting behavior

In all three studies we only examined the effect of IAT test difficulty on the relationship between IATs and direct attitude measures. The ultimate goal of using the IAT to predict outcomes should be to predict actual behavior and, accordingly, the final criterion for improving its predictive power should be to improve the relationship between IATs and behavioral measures. In our view, however, any means that leads to an increase in the systematic (i.e., true-score) variance of a measure has a good chance of increasing its predictive power for any outcome, whereas reducing its true-score variance is also likely to reduce its predictive power. We have shown that IATs with moderate difficulty have higher true-score variance, describe new approaches for using the concept of test difficulty to increase the true-score variance of IATs, and thus provide strategies for ultimately increasing the predictive power of IATs.

Conclusion

By introducing the test difficulty concept to IATs we have taken a new, test theoretical perspective to address a much debated issue in IAT research, namely the generally low predictive power of IATs. The evidence we presented suggests that this new perspective is indeed promising. We showed in two non-experimental studies (Study 1 and Study 2) that IAT test difficulty is a strong moderator of the IATs' predictive power, due to the fact that, as predicted by CTT, IATs of moderate difficulty (IAT effects close to zero) have higher true-score variance than IATs of extreme difficulty (IAT effects strongly deviating from zero). Moreover, we showed in an experimental study (Study 3) that IAT test difficulty can be used

to derive modifications of the IAT design to influence the test difficulty, true-score variance, and predictive power of IATs. Thus, we demonstrated not only the validity of the IAT test difficulty account, but also that IAT test difficulty provides new approaches to influence the true-score variance and the predictive power of IATs.

We related our findings to previous research and highlighted their broad implications for IAT research. These implications include, but are not limited to, the previously discussed examples, such as that the pursuit of strong ("reliable") IAT effects is counterproductive for the reliability of the IAT as a measure, and compromises the investigation and prediction of individual differences in correlational studies, or that established moderators suggest the development of IATs whose predictive power may even be diminished under certain conditions. We have also discussed what is known about factors that influence the test difficulty, true-score variance and predictive power of IATs, and have related these findings to the test difficulty account. In doing so, we have identified promising new directions in the context of the test difficulty account that might help to further address the problem of low predictive power of IATs, but which have not yet been explored. Accordingly, we provide an agenda for future research, including, for example, the investigation of whether the test difficulty, true-score variance, and predictive power of IATs can be manipulated via the attribute categories of IATs. Finally, we have discussed how the test difficulty account can be directly applied to one of the most prominent fields in IAT research, the field of implicit prejudice against and stereotypes of social groups. All in all, by applying the test difficulty account to IATs, we have provided a framework that will hopefully help to tackle the long-standing problem of low predictive power of IATs and stimulate new research in this direction.

References

Asparouhov, T., & Muthén, B. (2006). *Constructing covariates in multilevel regression*. Mplus Web Notes: No. 11. http://www.statmodel.com

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation* [Version 3]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.310.3903

Blair, I. V. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. https://doi.org/10.1207/s15327957pspr0603_8

Blanton, H., Jaccard, J., & Burrows, C. N. (2015). Implications of the Implicit Association Test D-Transformation for Psychological Assessment. *Assessment*, *22*(4), 429–440. https://doi.org/10.1177/1073191114551382

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192–212. https://doi.org/10.1016/j.jesp.2005.07.003

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT. *The Journal of Applied Psychology*, *94*(3), 567-582. https://doi.org/10.1037/a0014665

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*(6), 977–997. https://doi.org/10.1002/ejsp.487

Breitsohl, H. (2019). Beyond ANOVA: An Introduction to Structural Equation Models for Experimental Designs. *Organizational Research Methods*, *22*(3), 649–677. https://doi.org/10.1177/1094428118754988

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

Curran, P. J., & Bauer, D. J. (2011). The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change. *Annual Review of Psychology*, *62*, 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011). *Frontiers in Psychology*, *6*, Article 2039. https://doi.org/10.3389/fpsyg.2015.02039

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989x.12.2.121

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. https://doi.org/10.1080/10463280600681248

Gast, A., & Rothermund, K. (2010). When old and frail is not the same: Dissociating category and stimulus effects in four implicit attitude measurement methods. *Quarterly Journal of Experimental Psychology*, *63*(3), 479–498. https://doi.org/10.1080/17470210903049963

Gawronski, B., De Houwer, J., & Sherman, J. (2020). Twenty-five years of research using implicit measures. *Social Cognition*, *38*, 1–25.

Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung* (2nd ed.). VS Verlag für Sozialwissenschaften / Springer Fachmedien Wiesbaden GmbH Wiesbaden. https://doi.org/10.1007/978-3-531-93192-0

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). *Texts in statistical science series*. CRC Press Taylor and Francis Group.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology*, *40*(3), 357–365. https://doi.org/10.1016/j.jesp.2003.07.002

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., . . . Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, *54*(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

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 Social Psychology*, *74*(6), 1464–1480. https://doi.org/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. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. (2019). *bain: Bayes factors for Informative Hypotheses. R package version 0.2. 1* (Version 0.2.3) [Computer software]. https://CRAN.R-project.org/package=bain

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology, 71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Gulliksen, H. (1945). The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, *10*(2), 79–91. https://doi.org/10.1007/bf02288877

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. The Guilford Press.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Hussey, I., Hughes, S., & Nosek, B. A. (2018). The implicit and explicit Attitudes, Identities, and Individual Differences (AIID) Dataset. https://doi.org/10.17605/OSF.IO/PCJWF

Irving, L. H., & Smith, C. T. (2020). Measure what you are trying to predict: Applying the correspondence principle to the Implicit Association Test. *Journal of Experimental Social Psychology*, *86*, 1–14. https://doi.org/10.1016/j.jesp.2019.103898

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. https://doi.org/10.1037/0022-3514.91.1.16

Koch, T., Schultze, M., Jeon, M., Nussbeck, F. W., Praetorius, A.-K., & Eid, M. (2016). A Cross-Classified CFA-MTMM Model for Structurally Different and Nonindependent Interchangeable Methods. *Multivariate Behavioral Research*, *51*(1), 67–85. https://doi.org/10.1080/00273171.2015.1101367

Kurdi, B., & Banaji, M. R. (2017). Reports of the Death of the Individual Difference Approach to Implicit Social Cognition May Be Greatly Exaggerated: A Commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, *28*(4), 281–287. https://doi.org/10.1080/1047840X.2017.1373555

Kurdi, B., & Banaji, M. R. (2019). *Relationship between the Implicit Association Test and explicit measures of intergroup cognition: Data from the meta-analysis by Kurdi et al. (2018).* https://doi.org/10.31234/osf.io/vpcx8

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing Implicit Prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. https://doi.org/10.1111/spc3.12023

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes. *Multivariate Behavioral Research*, *39*(4), 653–686. https://doi.org/10.1207/s15327906mbr3904_4

Lord, F. M. (1953). The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, *13*(4), 517–549. https://doi.org/10.1177/001316445301300401

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

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*, 2483. https://doi.org/10.3389/fpsyg.2019.02483

Meissner, F., & Rothermund, K. (in press). Increasing the validity of implicit measures: New solutions for assessment, conceptualization, and action explanation. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit bias and racism.* Cambridge University Press.

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. https://doi.org/10.1037/a0030734

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Experimental Psychology*, *48*(2), 107–122. https://doi.org/10.1026//0949-3946.48.2.107

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of Personality and Social Psychology*, *85*(6), 1180–1192. https://doi.org/10.1037/0022-3514.85.6.1180

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.).

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nunnally, J. (1970). *Introduction to psychological measurement*. McGraw-Hill. http://worldcatlibraries.org/wcpa/oclc/76488

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3. ed.). *McGraw-Hill series in psychology*. McGraw-Hill.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 255–277). The Guilford Press.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of Mean and Covariance Structure Analysis: Integrating Correlational and Experimental Approaches. *Organizational Research Methods*, *7*(1), 27–65. https://doi.org/10.1177/1094428103259554

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel Structural Equation Models for Assessing Moderation Within and Across Levels of Analysis. *Psychological Methods*, *21*(2), 189–205. https://doi.org/10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A General Multilevel SEM Framework for Assessing Multilevel Mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

R Core Team (2021). R: A language and environment for statistical computing. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology*, *62*(1), 84–98. https://doi.org/10.1080/17470210701822975

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

Steffens, M. C., & Plewe, I. (2001). Items' Cross-Category Associations as a Confounding Factor in the Implicit Association Test. *Experimental Psychology*, *48*(2), 123–134. https://doi.org/10.1026//0949-3946.48.2.123

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60. https://psycnet.apa.org/record/1990-11390-001

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Urban, M., Koch, T., & Rothermund, K. (2024). The Implicit Association Test and its Difficulty(ies): Introducing the Test Difficulty Concept to Increase the True-Score Variance and, Consequently, the Predictive Power of Implicit Association Tests. https://osf.io/ex9ar/

Van Lissa, C. J., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., & Hoijtink, H. (2021). Teacher's Corner: Evaluating Informative Hypotheses Using the Bayes Factor in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(2), 292–301. https://doi.org/10.1080/10705511.2020.1745644

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian Approach to More Stable Estimates of Group-Level Effects in Contextual Studies. *Multivariate Behavioral Research, 50*(6), 688–705. https://doi.org/10.1080/00273171.2015.1090899