The Implicit Association Test paradigm

Rosseel, Laura

Published in:
Research Methods in Language Attitudes

Publication date:
2022

License:
Unspecified

Rosseel, L. (2022). The Implicit Association Test paradigm. In R. Kircher, & L. Zipp (Eds.), *Research Methods in Language Attitudes* (pp. 250-267). Cambridge University Press.

16 The Implicit Association Test Paradigm

Laura Rosseel

16.1 Introduction

The Implicit Association Test (IAT) is a reaction time-based categorisation task that aims to capture to what extent participants automatically associate an attitude object and a certain type of evaluation. The IAT was introduced some twenty years ago as a novel attitude measure in social psychology (Greenwald et al. 1998). In this field, the method has been used to study a wide range of topics such as racial bias (e.g. Greenwald et al. 1998), advertising and marketing (Maison et al. 2004; Gregg and Klymowsky 2013), self-esteem (e.g. McKay et al. 2010), gender bias (e.g. Robinson et al. 2005; Steele and Ambady 2006), and addiction (e.g. Houben and Wiers 2006; Lindgren et al. 2013). More recently the IAT paradigm has gained ground as a measure for language attitudes in the social and behavioural sciences. Linguists have employed the IAT to study the evaluation of language varieties and variants (e.g. Babel 2010; Redinger 2010; Pantos and Perkins 2012; Lee 2015; Loudermilk 2015; Leinonen 2016; Álvarez-Mosquera 2017; McKenzie and Carrie 2018), but also to investigate their indexicality and social meanings (Campbell-Kibler 2012, 2013b; Llamas et al. 2016; Hilton et al. 2016; Nilsson et al. 2019). The method has even been applied to measure salience (Leinonen 2016).

16.1.1 Implicitness

It is important to briefly consider the type of psychological construct measured with the IAT. The method is often said to capture implicit attitudes, but this concept of implicitness and the use of the term is rather controversial – at least in social psychology (see Chapter 1). *Implicit* is both used to describe a type of method (those that do not require introspection), as well as the construct to be measured (attitudes that fall outside a person's control and of which they are unaware). Yet, some authors prefer to use the term *indirect* to refer to methods that require no introspection. Others, like Gawronski and De Houwer (2014: 284) specify the term *implicit measure* by reserving it for cases where 'the to-bemeasured psychological attribute influences participants' responses on the task in an automatic fashion', whereby automatic is defined as 'unintentional, resourceindependent, unconscious, or uncontrollable'. For a more thorough, yet accessible discussion of the interpretation of the term *implicit* in social psychological attitudes research, refer to Pantos (2019). A further complication is how these concepts relate to the interpretations of different types of language attitudes (e.g. overt vs. covert attitudes, see Chapter 1) and language attitudes measures in the linguistic tradition. This is a pressing question that deserves further attention in future work, but is beyond the scope of this chapter. Interesting discussions in this context can be found in work like Adams (2019), Pantos (2019), Pharao and Kristiansen (2019), and Rosseel and Grondelaers (2019). In particular, the suggestion raised in Campbell-Kibler (2012) to further look into the similarities between the psychological concept of implicit attitudes and the sociolinguistic construct of indexical meanings as well as the underlying theoretical models seems highly relevant. Notwithstanding this lack of research integrating sociolinguistic and social psychological theorising on attitudes, it may be clear that the IAT fits well in the collection of chapters on indirect language attitudes methods in the sense that participants do not self-report their language attitudes during an IAT.

In this chapter, the term *implicit* is used for both the type of attitudes measured by the IAT as well as the method itself, as is customary in the research tradition the IAT paradigm emerged from, and to avoid confusion with the linguistic definition of indirect language attitudes measures.

16.1.2 How the IAT Works

The IAT measures the relative association between an attitude object and an evaluative dimension (for a discussion of the main evaluative dimensions of language attitudes, see Chapter 1). The attitude object is referred to as the *(target) category*, while the evaluation is termed the *attribute (category)*. So in an experiment that aims to measure language attitudes towards a linguistic phenomenon (e.g. language varieties or a linguistic variable) in terms of the status dimension, the linguistic phenomenon would be the target category and the status dimension the attribute. Note that both target and attribute need to be binary in an IAT. In this example, the linguistic phenomenon could consist of two language varieties or two different realisations of a linguistic variable and the attribute category would be 'status' versus 'no status'. This means that the attitudes measured using the IAT are always relative in nature. The implications of this binary structure are discussed in Section 16.3.

A standard IAT consists of a series of trials. These trials are categorisation tasks in which participants sort four types of stimuli representing the two target categories and the two attribute categories. These stimuli have to be highly representative of their respective category, so they are easily categorisable. For example, in a study on the relative status of two language varieties, variety A and variety B, brief sound clips representing each variety could function as target stimuli and adjectives expressing status or lack thereof (e.g. *highly educated, important, rich, uneducated, irrelevant, poor*) as attribute stimuli. To sort stimuli into their respective categories, participants have two response keys at their disposal.

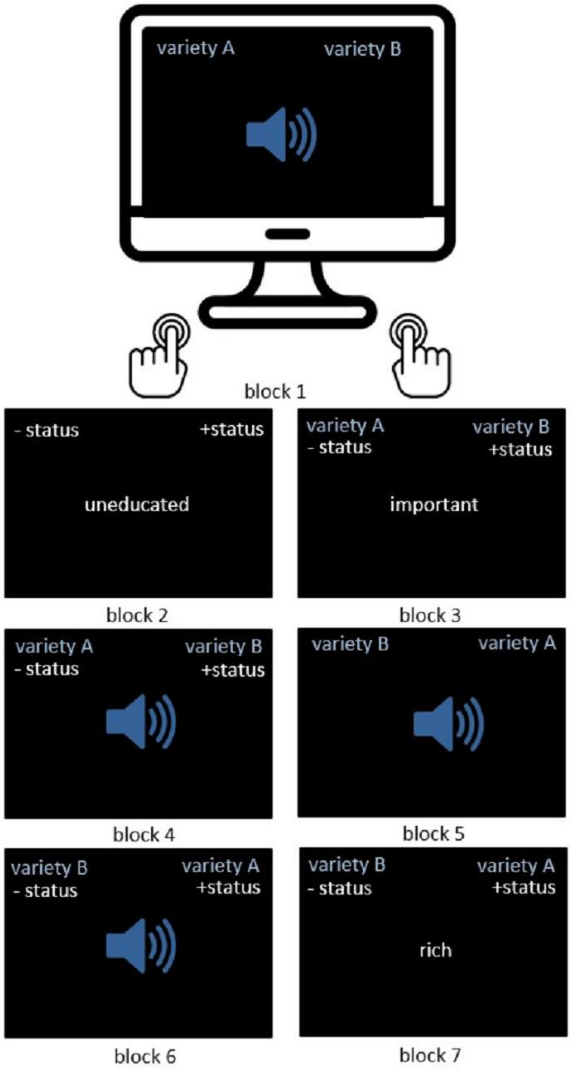

Figure 16.1 *Screenshots with an example of a trial from each block of an IAT*

In each trial, one stimulus has to be categorised using the two response keys. Examples of trials can be found in Figure 16.1. The stimulus appears in the middle of the screen after which the participant is asked to assign it to the correct category as quickly as they can, but at the same time as accurately as possible. Figure 16.1 also shows that in addition to the stimulus, there are two labels in the top corners of the screen. These labels indicate the meaning of the response keys, so the participant does not need to memorise how the target and/or attribute categories are mapped onto the response keys.

Table 16.1 – Schematic overview of the structure of an IAT measuring status associations with two language varieties

Block number	Type of block	Left response key	Right response key	Typical number of trials
1	Practice	Variety A	Variety B	20
2	Practice	-status	+status	20
3	Experimental	Variety A -status	Variety B +status	20
4	Experimental	Variety A -status	Variety B +status	40
5	Practice	Variety B	Variety A	40
6	Experimental	Variety B -status	Variety A +status	20
7	Experimental	Variety B -status	Variety A +status	40

The trials of an IAT are divided into seven blocks that differ in the types of stimuli that have to be categorised and in the mappings of the categories onto the response keys. This structure is summarised in Table 16.1 that builds on the example of an experiment measuring status associations with variety A and B. Three of the seven blocks (blocks 1, 2, and 5) are practice blocks and aim to familiarise participants with the stimuli and categorisation task. The four remaining blocks (3, 4, 6, and 7) are experimental blocks that serve as the basis to detect the IAT effect.

In the first practice block (i.e. block 1), participants practise categorising target stimuli. In the example in Figure 16.1, these consist of short sound samples representing variety A and B. The response keys in these blocks correspond to one target category each. Variety A is mapped onto the left key, variety B onto the right key (see Table 16.1 and Figure 16.1). This block is followed by another practice block (i.e. block 2), but this time the attribute stimuli are practised and the response keys now correspond to the two attribute categories (i.e. _status for the left key and +status for the right key in the example in Table 16.1). These practice blocks usually consist of twenty trials each.

Blocks 3 and 4 are the first set of experimental blocks. They combine the two tasks practised in blocks 1 and 2. Respondents now have to categorise a mix of target and attribute stimuli. However, now there are four types of stimuli for only two response keys. This means that two categories will be mapped onto the same response key. In the example, the left key is used for either stimuli representing variety A or stimuli representing -status and the right key for stimuli representing variety B or stimuli

representing +status. Note that for each type of stimulus only the relevant category applies: When participants get a target stimulus, only the target categories matter and the only significance of the keys is 'variety A' or 'variety B', and likewise when an attribute stimulus is presented. Together, these experimental blocks contain around sixty trials.

After this first set of combined blocks, participants go back to a practice block (i.e. block 5) with target discrimination. The reason why they have another practice round is that the mapping of categories onto the response keys has been swapped around: The left key now stands for variety B, while the right key is used to categorise stimuli representing variety A. It is recommended to have around forty trials in this practice block.

The final two blocks (i.e. blocks 6 and 7) are again combined blocks that require participants to categorise a mix of semi-randomly presented target and attribute stimuli. In these two blocks, the response key mappings from block 5 are retained, while the mappings for the attribute categories do not change throughout the experiment. This results in a reverse key mapping compared to blocks 3 and 4. So in the example the key mappings in the final two blocks are 'variety B / -status' for the left key and 'variety A / +status' for the right key.

This reversed response key mapping between the two sets of experimental blocks is crucial for the IAT. Participants will find it easier to categorise stimuli when the categories mapped onto the response keys are congruent with their attitudes. By contrast, they will find it more difficult when the category mapping is incongruent with their own attitudes. For example for the IAT design in Table 16.1, a participant who associates variety A more strongly with status than variety B will find it easier to categorise stimuli in the first set of experimental blocks (i.e. blocks 3 and 4) than in the second set (i.e. blocks 6 and 7), because in the former, variety A and +status are mapped onto the same response key. This difference in ease is reflected in a difference in reaction times: Respondents will be faster to categorise stimuli in blocks where the response key mappings are congruent with their attitudes, but slower when their attitudes are incongruent with the response key mappings. Hence, through a comparison of reaction times between the two sets of experimental blocks, the IAT provides information on which target and attribute categories participants associate more strongly.

16.2 Research Planning and Design

This section discusses stimulus selection, stimulus presentation and error feedback, programming possibilities, the practical circumstances for conducting an IAT experiment and ethical considerations.

Stimulus selection is a crucial aspect of designing a successful IAT experiment. A first point in that respect is the number of stimuli typically used. There is a lot of variability to be found in the literature on this point. Often, five to six stimuli per category are included, for a total of twenty to twenty-four stimuli per experiment. Nosek et al. (2005), however, showed that as long as more than two stimuli are used, the results are not affected significantly by the number of stimuli. What does matter, though, is the quality of the stimuli. It is important to ensure the stimuli are representative of their category: They have to be good exemplars of whatever attitude object they represent as they have a considerable influence on the outcome of the experiment. Mitchell et al. (2003), for instance, found that racial attitudes (in this case attitudes towards dark- and light-skinned people) changed depending on whether the stimulus set contained popular and well-liked black people for the category of dark-skinned people, and unpopular and disliked white people for the category of light-skinned people compared to an experiment where it was the other way around. Another reason why the stimuli have to be chosen carefully is that they need to be easy to categorise for participants, so they can do it quickly. This helps to ensure that automatic associations are measured with the IAT, which is also why the stimuli have to be as short as possible. The same care has to be taken when choosing the labels for the target and attribute categories that are displayed in the top corners of the screen to help participants remember the meaning of the response keys (see Figure 16.1). Multiple studies have shown that the choice of labels is just as decisive for the outcome of the experiment as the choice of

the stimuli (e.g. De Houwer 2001). The fact that the construal of the attitude object through its naming is an important factor for attitudes is also well-known in linguistic attitudes research where studies have pointed to the different outcomes of experiments measuring language attitudes based on variety labels versus sound clips (e.g. Bishop et al. 2005; Coupland and Bishop 2007; Kristiansen 2010b; Grondelaers and Kristiansen 2013).

As regards stimulus presentation, one has to decide on the intertrial interval. There is quite some variation between studies, with most studies leaving around 250 ms between a participant's response and the presentation of the next stimulus (Teige-Mocigemba et al. 2010). When participants miscategorise a stimulus, they receive error feedback usually in the form of a red cross in the middle of the screen. They then have to correct themselves by pressing the correct key before moving on to the next trial. Note that not all versions of the IAT use error feedback (see Section 16.3).

The next point relates to programming an IAT experiment. The procedure that gives the most freedom to control and adjust all aspects of an experiment is to programme it from scratch in whatever programming language one is versed in. There are alternatives however: One option is to use open source software like Psychopy (Peirce et al. 2019) which gives a reasonable amount of freedom to tailor the experiment to one's needs. Given the popularity of the IAT paradigm, it is also integrated in commercial software like E-prime (Psychology Software Tools 2016), which is often user-friendly and requires little programming skills, but offers less freedom to tailor certain aspects of the experiment.

Another practical issue to consider is the circumstances under which the IAT is taken. IAT experiments have been conducted both in lab conditions and online. The former option is perhaps more common and has the advantage of optimal control over the experimental setting, which is desirable for reaction time-based tests where the slightest distraction may impact participants' performance. However, it is time-consuming and requires considerable resources to conduct lab-based research especially when large samples of respondents are needed. Online versions of the IAT allow the researcher to cost-efficiently collect more data in a shorter amount of time, and perhaps to reach participant groups that are harder to bring into the lab (Friese et al. 2007; Glashouwer et al. 2013). The pay-off is that the quality of the data may be less than desirable, as there is no control over the circumstances in which participants take the test. Sometimes respondents are asked whether they were interrupted at any point during the online IAT as a control question to filter out participants who were distracted. This trade-off between highly controlled experiments, and cost-effective and fast data collection is of course not unique to the IAT paradigm, and researchers should weigh up what the best solution is for their study. More generally, it should be noted that the characteristics of a standard IAT design reported above are not set in stone: Considerable variation can be found in the implementation of various design features.

Finally, as with all experiments involving human participants, it is crucial to follow ethical guidelines carefully in the context of an IAT study. Researchers must make sure to obtain informed consent from respondents and provide adequate debriefing. It is also strongly recommended to seek institutional ethics clearance before embarking on an IAT study (see also Project Implicit 2011 for further ethical considerations regarding the IAT).

16.3 Strengths and Limitations

16.3.1 Strengths

16.3.1.1 Psychometric Qualities

The IAT paradigm has been used extensively in social psychological research and there are multiple reasons why the method has proven to be appealing to many scientists. A first of those is its psychometric qualities, which refers to the method's validity and reliability. A valid method is one that measures what it set out to measure. For an implicit attitude measure to be valid, the effects it captures should reflect variations in the targeted attitudes. Reliability refers to a method's consistency: A reliable method should produce the same results under the same circumstances. The validity and

reliability of a method depend on the specifics of a study (which attitudes are measured and what procedural choices have been made). Given that the IAT is used to assess a wide variety of attitudes and that there is considerable variation in its procedural implementation, one has to be careful with generalisations about its psychometric qualities (Lane et al. 2007). However, considering evidence from a large range of studies, the IAT is deemed to have good psychometric qualities (Gawronski et al. 2011). Generally, psychologists assume that the IAT effects are caused by the attitudes the measure tries to capture (De Houwer et al. 2009). Hence, it is seen as a valid measure. Reliability for the IAT in its traditional form is also satisfactory (internal consistency is usually situated between .70 and .90, see Gawronski and Hahn 2019).

16.3.1.2 Flexibility

In addition to good psychometric properties, the IAT paradigm is also popular for its flexibility with regard to stimulus modality and the types of evaluations it can measure. The IAT allows for any stimulus modality, for example written (nonsense) words, pictures, or sounds (e.g. Vande Kamp 2002; Lane et al. 2007). Especially the latter option makes it an appealing method for researchers interested in measuring language attitudes. Although many psychological studies limit the attribute category to valence, the IAT allows any type of evaluation to feature as the attribute concept, as long as confounds with the target concept are avoided. This is again particularly interesting for researchers of language attitudes who are typically interested in a more diverse range of attitudinal dimensions such as status and solidarity (see Chapter 1). Note though that only one attitudinal dimension can be included per IAT experiment.

The flexibility on the level of the stimuli has its bounds, however. Given that speed is often cited as an important component of automaticity, a considerable restriction is that the stimuli need to be as short as possible to maintain the implicit nature of the method: The longer the stimuli, the more time participants have to process them and the less automatic their reactions become. Typically, vocal stimuli used in linguistic IAT studies are short words of up to two syllables (ca. 600 ms, e.g. Rosseel et al. 2018, but see, for instance, Pantos and Perkins 2012 for auditory stimuli of up to twelve words, ca. twenty-four syllables). This limits the type of linguistic phenomena that can be studied using the IAT. It will, for instance, be hard to include stimuli that contain certain syntactic or prosodic features that require stimuli to be longer than just a few syllables. The length restriction on the stimuli in the IAT comes with another consideration: It leaves the stimuli completely decontextualised. While this may present opportunities to study the social meaning of linguistic features in isolation in a carefully controlled experimental setting, it comes with the disadvantage that it is hard to study language attitudes in a more ecologically valid way. It also makes it challenging to use the method to tap into the mediation of attitudes by situational, social, and linguistic contextual information. Although previous studies have attempted to contextualise the IAT (e.g. Rosseel et al. 2019), it seems anything but straightforward to do so.

16.3.1.3 Faking

A rather obvious advantage of the IAT paradigm is that it allows to measure implicit attitudes. It was explained above that, as an implicit attitude measure, the responses in the IAT's categorisation tasks should be influenced by participants' attitudes in an unintentional way (Gawronski and De Houwer 2014: 284). For this to be true, participants should not be able to influence the outcome of the IAT intentionally. Consequently, a considerable number of studies have looked into the extent to which respondents can 'fake' an IAT. Although it is not impossible, the IAT has been shown to be rather fake-proof (Teige-Mocigemba et al. 2010). If faking is successful, it is usually only the case in participants with prior experience with the method (Steffens 2004; Fiedler and Bluemke 2005) or in respondents to whom it has been explained how they can fake it (Fiedler and Bluemke 2005). Hence, it is advisable not to include participants who have prior familiarity with the method in an IAT study. Another way to avoid strategic control of the responses is to opt for an alternative but closely related method to the

IAT, the Implicit Association Procedure (see Schnabel et al. 2006). This method relies on the use of a joystick to categorise stimuli rather than on response keys and is believed to be more immune to faking (Teige-Mocigemba et al. 2010).

16.3.2 Limitations

16.3.2.1 Labelling Target and Attribute Categories

Certain features of the IAT can make its use challenging or unsuitable for certain purposes. A first disadvantage is the method's requirement to label the target and attribute categories. As explained above, these labels are used to communicate the mapping of the target and attribute categories onto the response keys (see Figure 16.1). Choosing a suitable label may prove difficult, for example, in a study that aims to measure the social meaning of a linguistic feature for which laypeople have no terminology. And if a label is available, it may be unclear what precisely non-linguists mean when using that label and whether they all mean the same. This is a particularly important issue, given that the choice of how to label an attitude object may impact the attitudes measured, as discussed in Section 16.2. Hence, careful pretesting and preparatory research in the speech community under study is primordial. If the label is chosen well, it may help to ensure that all participants identify the linguistic phenomenon under study as it was intended by the researcher. As such, the use of labelling in the IAT circumvents an issue that, for instance, the matchedguise technique is sometimes criticised for (Garrett 2010; see also Chapter 12). Note also that there may be creative solutions in cases where no obvious label is available, such as the labelling of supposed speakers (e.g. person A vs. person B, or John vs. Peter), or by asking the participant to provide their own labels.

16.3.2.2 Binary Structure

A second potential limitation of the IAT is its binary structure. As explained in Section 16.2, this requires researchers to work with binary attitude objects and evaluative dimensions. This can of course be inconvenient for studies focusing on single attitude objects (e.g. a single language variety) or on more than two attitude objects (see Section 16.6 for an example). Yet, for both situations, solutions are available. Alternative versions of the IAT have been developed to accommodate for single attitude objects (Single Category IAT, see Karpinski and Steinman 2006; or Single Target IAT, see Wigboldus et al. 2004) or a single attribute category (Single Attribute IAT, see Penke et al. 2006). In these types of IATs, it is advisable to use target and attribute stimuli of the same modality in order to avoid recoding (Gawronski et al. 2011). What could happen is that participants use the modality of the stimulus as the criterion to categorise it rather that the intended target and attribute concepts. In that case the IAT is not measuring what it intended to measure anymore. However, using target and attribute stimuli of the same modality may pose problems in the case of auditory linguistic stimuli. An example would be a Single Target IAT used to measure positive/negative attitudes towards a certain accent represented by auditory stimuli. To avoid recoding in this example, the attribute stimuli would also have to be auditory. The question then is which accent to use for the attribute stimuli. If the same accent is used for the target and attribute stimuli, the distinction between the two is not clear enough anymore for the IAT to function properly. Yet, if another accent is used for the attribute stimuli, a confound would be introduced into the design. One could argue that a neutral accent can be used for the attribute stimuli, but it may obviously be questioned whether there is such a thing as a neutral accent.

A solution for studies that aim to measure attitudes towards more than two attitude objects is to develop multiple IATs that each include a different pairing of the attitude objects. Such multiplication of the number of experiments is not always desirable: Although studies with successful within-subject designs with multiple IATs have been reported, there is a risk of practice effects or fatigue in participants' performance (Fiedler and Bluemke 2005; Gawronski et al. 2011; Bar-Anan and Nosek 2014). Finally, note that the binary structure of the standard IAT entails that the method can only measure relative attitudes, never absolute ones. This should however not necessarily be regarded

as a disadvantage. Language attitudes are inherently relative: The use of a linguistic feature is always judged in comparison to alternative features that could have been used in the same place. And if that is the case, the binary structure of the IAT forces the researcher to specify what the second component of the comparison is, hence allowing for a more controlled experimental design and enhanced validity.

16.3.2.3 Block Order Effects

A third problematic aspect of the IAT paradigm is also related to its structure: The method suffers from block order effects. It has been shown that the IAT effect tends to be larger if the congruent experimental block (i.e. where the response key mappings align with the participant's attitudes) precedes the incongruent experimental block (Teige-Mocigemba et al. 2010). In many studies it is of course unknown in advance what a participant's attitudes will be and hence what the congruent block will be for them. One way of dealing with this order effect is to counterbalance block order between participants. Note that a disadvantage of this solution is that there will be more error variance in the results which may decrease the correlation with other attitudes measures (De Houwer et al. 2015). Alternatively, or additionally, the block order effect can be diminished by increasing the number of stimuli in block 5 from the traditional 20 to 40 (Nosek et al. 2005). This gives participants more time to get used to the swapped key mappings. Note that this solution has already been implemented in the IAT design as described in Section 16.1. A third way of tackling block order effects is to opt for alternative versions of the IAT designed to deal with this issue by getting rid of the block structure (e.g. the Single Block IAT, see Teige-Mocigemba et al. 2008; the Recoding Free IAT, see Rothermund et al. 2009).

16.3.2.4 Extra-Personal Associations

A final challenge associated with the IAT to be discussed here is the type of associations measured by the method. The standard IAT as described in Section 16.1 has been criticised for measuring extra-personal associations rather than personal ones (note that this distinction is debated, see Gawronski et al. 2008; De Houwer et al. 2009). These extra-personal associations refer to associations which are picked up because they are present in society, but which the individual may not endorse. For example, an IAT comparing vegetables to candy bars may return positive associations with the former even for participants who prefer candy bars to vegetables. The reason for this is that these participants live in a society where they are regularly told that vegetables are good for them and it is this knowledge the IAT taps into. To circumvent this problem, Olson and Fazio (2004) introduced the Personalised Implicit Association Test (P-IAT) as a new variant of the traditional IAT. The two main elements that set the P-IAT apart are the omission of error feedback after an incorrectly categorised stimulus and the use of personalised attribute labels (e.g. 'I like' vs. 'I don't like' instead of 'positive' vs. 'negative'). This version of the IAT is used in the case study presented in Section 16.6.

16.4 Data Analysis and Interpretation

Data recorded during the IAT, that is, response latencies and accuracy, are typically analysed using one of three different procedures that aim to compare participants' performance between the two sets of experimental blocks. In the traditional approach (Greenwald et al. 1998), extreme latencies are first dealt with by recoding those that are too fast (i.e. <300 ms) to 300 ms and those that are too slow (i.e. >3,000 ms) to 3,000 ms. These are typically responses where participants accidentally pressed the response key too soon in anticipation (<300 ms) or were distracted (>3,000 ms). Next, reaction times are log-transformed and the difference is calculated between the mean reaction time in experimental blocks 7 and 4. A second approach is to use a scoring algorithm to calculate what is referred to as measure *D* (Greenwald et al. 2003). Although different versions of the scoring algorithm exist, the one recommended by Greenwald et al. (2003) is one that calculates the average of the difference between the mean reaction times of the two sets of experimental blocks (mean block 6 minus mean block 3,

and mean block 7 minus mean block 4) divided by the inclusive standard deviation of each difference score. Note that before calculating the average difference, all trials are excluded that have reaction times over 10,000 ms. This is done to deal with the distribution of latency data, which often has misshapen tails. Furthermore, all participants with more than 10 per cent of their responses under 300 ms are removed from the data. These participants are likely responding too fast to properly classify the stimuli. Greenwald et al. (2003) present these exclusion criteria as a minimum and recommend considering further exclusions based on fast/slow responses and error rates in laboratory studies. For those using R statistical programming software to analyse their IAT data, the IAT package (Martin 2015) can be used to tidy up IAT output and calculate *D* scores. As an alternative to the traditional approach and the *D* score, some researchers use mathematical approaches like multinomial modelling to analyse IAT data (see Sherman et al. 2010 for a review).

When using the scoring algorithms described above, it is crucial to keep in mind that they present difference scores and that the IAT is a relative attitude measure of which the outcome cannot be interpreted in an absolute manner. The difference scores express the association between the four target and attribute categories relative to one another. Going back to the example IAT used in Section 16.1, one could never claim that 'participants feel variety B lacks status'. All one could deduce from a difference score in this case would be that participants perceive variety B as less strongly associated with status in comparison to variety A. No absolute claims could be made regarding the status of either variety. Likewise, a difference score of zero cannot be interpreted as indicative of a neutral attitude. For a further critical appraisal of the use of difference scores in the context of the IAT, see Teige-Mocigemba et al. (2010).

16.5 Further Important Considerations

Before turning to an example of a linguistic IAT study, this section briefly touches on two issues researchers should consider before embarking on any IAT study. First, the IAT paradigm is not the most suited methodology for exploratory studies of language attitudes. The IAT's structure, with a predefined attitude object and evaluative dimension, lends itself better to testing specific hypotheses. It is important to remember that the standard IAT only allows the inclusion of a binary target object and just one evaluative dimension which also needs to be binary. Furthermore, practice effects and fatigue are risks that make repeated measures studies unappealing.

Second, it should not be underestimated that completing an IAT experiment is cognitively demanding for respondents. The computer-based task is rather complex and the test requires considerable concentration for a prolonged time. This means that it may not be suitable to use in all participant populations. However, to tackle this limitation, versions of the IAT have been developed that cater for specific respondent groups, like children as young as four years old (Cvencek et al. 2011; see also Chapter 15).

16.6 Case Study: Attitudes towards Language Variation in Belgian Dutch

To illustrate how the IAT paradigm can concretely be implemented in linguistic attitudes research, this section describes an experiment that was designed to study attitudes towards language variation in Belgian Dutch. The full study is reported in Rosseel et al. (2018). At the time this study was conducted, fairly little language attitudes research on Belgian Dutch had been carried out since the 1980s. Hence, the aim of the study was to contribute to painting a picture of the language attitudinal landscape in Flanders, the Dutchspeaking northern part of Belgium, and to a better understanding of processes of language variation and ongoing change there. The study also had a methodological objective, namely comparing the potential of the IAT paradigm with that of auditory affective priming as novel methods for measuring language attitudes (see also Chapter 20). In what follows, my focus is limited to the

practical aspects of the IAT experiment. For a discussion of the study's larger descriptive and methodological objectives, I refer the reader to Rosseel et al. (2018).

The case study set out to measure attitudes towards three varieties of Belgian Dutch, Standard Belgian Dutch (SBD) and two varieties of Colloquial Belgian Dutch (CBD), Antwerp-accented CBD and West-Flemish-accented CBD, in participants native to Antwerp and West-Flanders. Based on previous work (e.g. Speelman et al. 2013), it was hypothesised that both groups of participants would prefer SBD over the outgroup variety. Furthermore, we expected participants from Antwerp to prefer their own regional variety over SBD, while the opposite was expected for West-Flemish respondents. Given that we had specific hypotheses regarding the attitudes towards the three varieties (see Section 16.5) and that we were interested in automatic evaluations, the IAT paradigm was an appropriate methodological choice.

Data from 176 participants were included in the analyses. All participants were born and raised in Antwerp or West-Flanders and were still living there. In order to exclude the potential influence of age and social background on the outcome of the experiment, these factors were kept constant by opting for a sample of university students. The average age of participants was 20 (SD = 1.79), and 102 identified as male, while 74 identified as female.

In the study, we used a P-IAT to avoid measuring extra-personal associations (see Section 16.3.2.4). The target categories in the study were the three language varieties under study, namely Antwerp-accented CBD, West-Flemish-accented CBD, and SBD. Given that the P-IAT requires the attitude object to be binary, multiple P-IATs had to be designed to include all three varieties in the study (see Section 16.3.2.2). One P-IAT included the Antwerp variety compared to SBD, while the other P-IAT had the West-Flemish variety and SBD as its target categories. The target stimuli consisted of sound recordings of six individual disyllabic words (see Section 16.3.1.3). For every word a recording was made in each of the three varieties. The target stimuli were controlled for length (around 600 ms), frequency, familiarity, valence, and degree of colloquiality (for more details, see Rosseel et al. 2018). Note that because the stimuli are so short, the difference between the varieties is purely situated on the level of the accent. Because dialectal variation was not the focus of this study, lexical items were used that showed no variation on the level of the lexicon. The stimuli were produced by two native speakers who each provided stimuli in their respective native accent. Additionally, each speaker provided half of the SBD-accented stimuli. This practice resembles a combination of a matched- and verbal-guise design (see Chapters 12 and 13).

The labels used to refer to the varieties in the P-IAT (i.e. the labels in the top corners of the screen, see Figure 16.1) were Dutch translations of 'Antwerp accent', 'West-Flemish accent', and 'neutral accent'. Note that from a linguistic point of view, these may not be the most straightforward labels, but after carefully pretesting different options, these labels turned out to be the most intuitive for our non-linguist respondents (see Section 16.3.2.1). For the attribute category, that is, the attitudinal dimension, we opted for valence. We aimed to measure to what extent participants held positive or negative associations with the three varieties under study. The attribute stimuli were normed positively and negatively valenced real-life photographs (Spruyt et al. 2002). The labels used for the attribute categories were Dutch versions of 'I like' and 'I don't like'. Note that these are personalised labels in the sense that they refer to the personal opinion of the respondent, which is crucial for the P-IAT (see Section 16.3.2.4). A further aspect of the personalisation of the IATs used here is that error feedback was omitted for the attribute stimuli.

In order to tackle block order effects, block order was counterbalanced between participants (see Section 16.3.2.3). This means that for each of our two P-IATs (one for Antwerp-accented CBD vs. SBD and one for West-Flemish-accented CBD vs. SBD), two versions were developed with inverse block orders. Thus, the study contained four P-IATs in total. These are referred to as experiments A, B, C, and D in Table 16.2, which also shows the number of trials in each block. These numbers are similar to those described in the design of the standard IAT in Section 16.1.2, but notice that we increased the number of trials in blocks 3 and 6, and that we added an additional four trials (one for each of the

target and attribute categories, blocks Pre 3 and Pre 6) before each set of experimental blocks. These four extra trials were added to give participants an opportunity to get used to the combined categorisation task in the following two blocks. These trials were not taken into account in the analysis. Participant numbers were fairly evenly distributed over the four versions of the experiment (see Table 16.3).

The raw reaction times obtained in the P-IATs were converted to *D* scores as described in Section 16.4 using the IAT package in R (Martin 2015). Positive *D* scores indicate a preference for SBD over the regional variety in the sense that SBD is more strongly associated with positivity. A negative *D* score means a stronger association between positivity and the regional variety compared to SBD. *D* scores were analysed using linear regression modelling. Note that in the analyses presented below, we discuss the results of experiments A and B together as experiment AB and those of experiments C and D as experiment CD, given that we are not interested in analysing block order effects, but merely aim to counterbalance order to control for this effect (see Section 16.3.2.3).

Figure 16.2 displays the mean *D* scores in experiment AB (comparing Antwerp-accented CBD and SBD) and experiment CD (comparing West- Flemish-accented CBD and SBD) and summarises the main findings of the study. First, we notice that the mean *D* scores are positive in both experiments for each participant group. This means that both respondents from Antwerp and West-Flanders have more positive associations with the standard variety than either their own regional variety or that of the other group. In Rosseel et al. (2018) we interpret this as an indication of a strong standard language ideology.

Second, we see that these positive associations with SBD are not always equally strong. More concretely, Figure 16.2 shows an interaction between the language varieties in the P-IAT and the regional origin of the participants. For both groups of respondents, the association between SBD and positivity is stronger in the experiment that contains the other group's regional variety of CBD. Or put differently, language users' preference for SBD decreases when that variety is compared to their own regional variety of CBD than to another variety of CBD. This pattern could be interpreted as an indication of ingroup preference: When compared with the respondent's own variety, the positive associations with SBD are competing with positive associations with the participant's native variety of CBD. For a further discussion of these results, see Rosseel et al. (2018).

Finally, although our hypotheses were confirmed overall, we did not find participants from Antwerp to prefer their own variety over SBD. Crucially, this hypothesis was based on a priming study. In Rosseel et al. (2018) we propose a number of methodological explanations for why the IAT results slightly diverge from the outcome of a priming experiment based on the same stimulus materials. As discussed in Section 16.2, attitudes measured with an IAT are based on both labels and stimuli. In a priming experiment, no labels are used. Hence, we suggested that the explicit presence of a label referring to SBD may have led to the measurement of attitudes in which the dominant standard language ideology in Flanders shines through more than in the priming experiment, where explicit labelling of varieties was absent.

Rosseel, L. (2022). The Implicit Association Test paradigm. In R. Kircher, & L. Zipp (Eds.), *Research Methods in Language Attitudes* (pp. 250-267). Cambridge University Press.

Table 16.2. *P-IAT structure experiments A, B, C, and D (Antw = Antwerp-accented CBD; WFl = West-Flemish-accented CBD)*

		Experiment							
		A		**B**		**C**		**D**	
block	**# trials**	**left key**	**right key**	**left key**	**right key**	**left key**	**right key**	**left key**	**right key**
1	20	Antw	SBD	SBD	Antw	WFl	SBD	SBD	WFl
2	20	I don't like	I like	I don't like	I like	I don't like	I like	I don't like	I like
Pre 3	4	Antw I don't like	SBD I like	SBD I don't like	Antw I like	WFl I don't like	SBD I like	SBD I don't like	WFl I like
3	40	Antw I don't like	SBD I like	SBD I don't like	Antw I like	WFl I don't like	SBD I like	SBD I don't like	WFl I like
4	40	Antw I don't like	SBD I like	SBD I don't like	Antw I like	WFl I don't like	SBD I like	SBD I don't like	WFl I like
5	40	SBD	Antw	Antw	SBD	SBD	WFl	WFl	SBD
Pre 6	4	SBD I don't like	Antw I like	Antw I don't like	SBD I like	SBD I don't like	WFl I like	WFl I don't like	SBD I like
6	40	SBD I don't like	Antw I like	Antw I don't like	SBD I like	SBD I don't like	WFl I like	WFl I don't like	SBD I like
7	40	SBD I don't like	Antw I like	Antw I don't like	SBD I like	SBD I don't like	WFl I like	WFl I don't like	SBD I like

Rosseel, L. (2022). The Implicit Association Test paradigm. In R. Kircher, & L. Zipp (Eds.), *Research Methods in Language Attitudes* (pp. 250-267). Cambridge University Press.

Table 16.3. *Distribution of participants over experimental design*

	Antwerp participants	West-Flemish participants
Experiment A	23	24
Experiment B	22	21
Experiment C	22	20
Experiment D	23	21

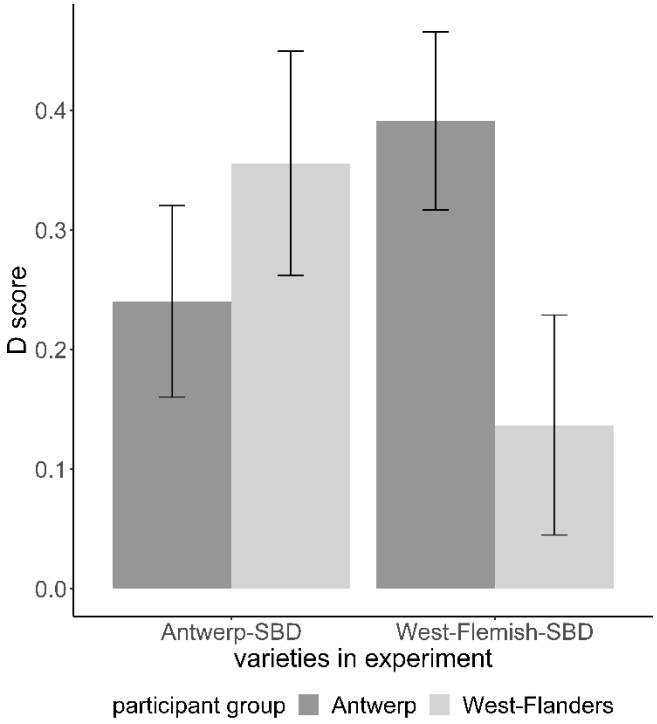

Figure 16.2 *Mean* D *scores per participant group for the P-IATs comparing associations with the two varieties of CBD vs. SBD**

** Positive* D *scores indicate a stronger association between the standard variety and positivity than between the regional varieties and positivity.*

Suggested further readings

Campbell-Kibler (2012); Gawronski et al. (2011); Greenwald et al. (1998); Pantos and Perkins (2012); Teige-Mocigemba et al. (2010)