Psychologie
Faculteit der Sociale Wetenschappen

Universiteit Leiden

Picture this: Comparing a picture-only

and word-picture race bias IAT

Name: Binnekamp, Jeffrey
Student number: 1403354
Date: 16/07/2019
Supervisor: Kret, M. E.
Second reader: TBA
Word count: 13.452
Cognitive Psychology
Thesis MSc Applied Cognitive Psychology

COMPARING A PICTURE-ONLY AND WORD-PICTURE RACE BIAS IAT

Abstract

Although they appear in a variety of fields of research, picture-only IATs (P-IATs) are currently rarely used. P-IATs lack the need for verbal processing, which make them useful for testing with children, analphabetic participants and non-human primates. In addition, P-IATs might be useful to circumvent requiring translations. Due to this potential use, we conducted an online questionnaire-based study using a within-subject design with a large sample size (N = 141) in which we compared a race-bias P-IAT with a word-picture race bias IAT (W-IAT). The P-IAT and W-IAT both yielded significantly distinct, moderately correlated IAT scores, with the W-IAT D-scores exceeding the P-IAT D-scores. Explicit bias was measured with the Symbolic Racism 2000 Scale (SRS), which was significantly correlated to the W-IAT, but not to the P-IAT. Neither IAT version was significantly correlated to reported outgroup familiarity. Instead, this was negatively correlated to the SRS. We discuss the potential reasons for these findings and provide recommendations for future research.

COMPARING A PICTURE-ONLY AND WORD-PICTURE RACE BIAS IAT

Table of contents

Introduction

Mental processes often involve associating one concept, category or subject to another. In doing so, our brains can access large amounts of information with relative ease. This is in part due to cognitive processes outside of our mere conscious thought. Unlike explicit attitudes, these processes are not easily accessed, as they are not accessible with introspection by definition (Nisbett & Wilson, 1977). In addition, implicit attitudes play a role in cognition distinct from explicit attitudes (Greenwald & Banaji, 1995), although the two are related to each other (Fazio, 1990; Hofmann, Gschwendner, Nosek & Schmitt, 2005a). Due to this, implicit attitude measurement has become an important part in predicting behaviour that cannot be accounted for by explicit measures (Nosek, Hawkings & Frazier, 2011).

Various measurements have been developed to measure the implicit associations that stem from these implicit attitudes (Nosek, Hawkins & Frazier, 2011); one of the most popular being the Implicit Association Test (IAT) introduced by Greenwald, McGhee and Schwartz (1998). This test uses a variety of stimuli, which are typically words combined with pictures. Only a handful of IAT studies have used pictures exclusively, despite the potential benefits such a design would bring. A picture-only IAT (P-IAT) can circumvent the need for participants to know the words used or even the need to be literate. This would allow testing with groups such as illiterate people, people who do not speak the language of any current IAT translation, people with certain mental disabilities, and even other species besides *Homo Sapiens*. Additionally, a P-IAT could circumvent some issues inherent in using translations of the IAT. The present article will therefore investigate the possibility of exclusively using pictures in an IAT. In order to do so, this paper will compare a P-IAT with a classical IAT that partially uses words (W-IAT) with a within-subject design with a large sample size (141 participants).

Before explaining how words or pictures have previously been used in an IAT, I will first provide a brief summary of what an IAT entails. After this explanation, the few previous studies that have used P-IATs will be discussed. This will be followed by a discussion of the benefits of such IATs and the theory behind the difference between using words and pictures. The final section of the introduction will discuss the goals and hypothesis of this study.

The Implicit Association Test

The IAT (Greenwald, McGhee & Schwartz, 1998) is a task designed to measure implicit associations through the difference in response times to a sorting task. The test works on the assumption that the presence of a strong implicit association will elicit a faster response time when sorting a category with a concept paired with an attribute. These associations are unconscious and might be present in people whilst they are unaware of it. For example, an IAT can compare the concepts of insects and flowers by pairing them with the attributes of pleasant or unpleasant. When someone has only ever had positive memories of receiving, smelling or seeing colourful flowers they might associate flowers with positive attributes. However, the same person is likely to dislike most insects, as insects are near-universally considered a nuisance, a pest or even dangerous. This would create strong and quick mental links and thereby negative associations with insects as a whole. Even if this person learns more about the ecological necessity of insects, they might still hold a strong negative association. In this case, the implicit association even exists in direct contrast to one's explicit opinions of insects. This insidiousness makes it hard to access the effect of these implicit associations, as they work outside of conscious thought. However, they might be at the root of attitudes and behaviour. This is why Greenwald, McGhee and Schwartz (1998) developed the IAT. In order to illustrate that the IAT can unearth these implicit associations, they conducted three experiments. In all three cases the expected differences in association speed were

significant. This indicates that the IAT can be used to test near-universal attitudes, such as expected attitudes for certain in-groups and out-groups, and for consciously denied implicit attitudes.

In the first of their three experiments, Greenwald, McGhee and Schwartz (1998) tested the aforementioned near-universal association of flowers (positive) versus insects (negative) and for musical instruments (positive) versus weapons (negative). They found that the compatible evaluative combination of concept and attribute (e.g. rose + pleasant or violin + pleasant) elicited a faster response time than an incompatible combination (e.g. wasp + pleasant or violin + unpleasant). This indicated the presence of an implicit association for these near-universal evaluative combinations. The second experiment tested the existence of expected differences in evaluative ethnic group associations. To do so, the IAT measured the attitudes of Japanese-Americans and Korean Americans to their own group and to the other group. It was expected that each group would be faster when their own ethnic group was paired with a positive attribute, and when the other ethnic group was paired with a negative attribute. The study confirmed their hypothesis, as the Japanese-Americans and Korean-Americans showed a positive implicit attitude to names from their own group (in-group bias), and a negative implicit attitude to names from the other group (out-group bias). In a third experiment, consciously denied implicit attitudes were tested with Caucasian subjects that described themselves as unprejudiced. In the ensuing IAT, the concepts were Afro-American names as compared to Caucasian names, and the attributes pleasant and unpleasant. Despite their denial of being prejudiced, the subjects still showed a positive bias to Caucasian names and a negative bias to Afro-American names.

After the initial study, many other studies started to use the IAT as a way to measure implicit attitudes for a wide variety of concepts. It has become a popular tool to measure implicit associations and has been used in more than 500 studies in the first 10 years after the

7

original study alone (Smith & Nosek, 2010). Although the original study used names as stimuli for the concepts, a review study as early as 2005 (Hoffmann, 2005) used a variety of different stimuli, including names, pictures of faces, acoustic stimuli and various types of words. Commonly, pictures are used for the concepts, but words for the attributes.

Previous P-IATs

To the best of my knowledge, only a few studies have used images exclusively, most of which were part of separate fields of research. One such IAT has been developed by Pieters, van der Vorst, Engels and Wiers (2010), who used a picture-only IAT to measure implicit associations on parental alcohol use in children. This IAT had alcohol/soft drinks and happy/angry faces as stimuli. A related IAT was developed by Palfai, Kanter and Tahaney (2016), who developed a pictorial alcohol IAT with university students with alcohol/water pictures and pictures with approach/avoidance behaviour. This IAT had adequate internal reliability. In a different field, Slabbick, Houwer and Kenhove (2011) developed an IAT for attitudes related to implicit motives, using attractive/unattractive pictures and pictures associated with or without the need for power. This was compared with non-IAT measures for implicit motives and with a verbal IAT. In another field of research, Thomas, Smith and Ball (2007) developed a fully pictorial measure of implicit associations for children with pictures of flowers/insects and with pictures of obese/thin adult females. They were successful in testing three to seven year olds with the version, which illustrated the benefits of using a picture-only IAT with young children.

This potential has been explored further in the line of research that uses race bias IATs with children and pre-schoolers. In this field, Newheiser and Olson (2012), exclusively used picture stimuli when creating their IATs for children from 7 to 11 years old. The study conducted two IATs: one to test implicit race bias, and one to test implicit social status bias.

8

For the concepts, the test used pictures of black and white faces in the race IAT, and pictures of 'rich' stimuli such as a sports car and 'poor' stimuli such as a dilapidated house for the social status IAT. For the attributes, the IATs used a variety of 'good' stimuli (which include a birthday present, flowers, puppies and ice cream) and 'bad' stimuli (which include a house on fire, a car crash, a spider and a snake). Notably, this study found significant results with both IATs. Newheiser et al. (2014) used this same race IAT and status IAT format with children from low status groups. Again, they found significant results with both IATs.

A related line of research with picture-only IATs is the more recent development of race bias IATs targeted to preschool children. At first this research focussed on creating IATs with audio clips synced with the appearance of words (Baron & Banaji, 2006; Cvencek, Meltzoff & Greenwald, 2011). However, this method still required verbal processing. This prompted Cvencek, Greenwald and Meltzoff (2011) to use a design that switches each trial between this words with audio variant and a variant with pictures only. Using this task as a basis, Qian et al. (2016) created a task that makes full use of pictures, which they call the Implicit Racial Bias Test (IRBT). This uses a smiling face icon or frowning face icon as stimuli in place of the word attributes seen in regular IATs. Thus, the IRBT uses only two images for attributes instead of two diverse groups of words and images. Unlike the regular IAT it requires the participant to press the face icons instead of regular keyboard keys. Using black and white faces for the concepts, the study found a strong cross-cultural implicit bias against other-race groups in 3-5 year olds from both China and Cameroon. The IRBT was later also used in a handful of studies (Qian, Heyman, Quinn, Fu & Lee, 2017; Qian et al., 2017a; Qian et al., 2017b; Setoh et al., 2019). Additionally, Rutland, Cameron, Milner and McGeorge (2005) and Steele, George, Williams and Tay (2018) used simple line drawings of a smile and frown for the attributes, but did not call their IATs an IRBT. Williams and Steele (2016) found that child IATs such as these had an internal consistency and test-retest

reliability consistent with that of adult IATs. In their study they used pictures for the concepts, and positive and negative line drawings (happy and sad faces) for their attributes.

Finally, a picture-only race IAT has been used by Van Berlo, Otten and Kret (in prep) with adults and children. This study found the same positive race bias for Dutch individuals and negative bias for Moroccan individuals for both age groups. Unlike previous studies, this IAT design was made for the explicit purpose of being used without relying on verbal processing at all, as it was designed for eventual use with non-human subjects. The aim of the study was to benefit comparative research between humans and bonobo apes by providing this non-verbal alternative IAT.

Advantages of P-IATs

The non-reliance of a picture-only IAT on verbal processing can provide a number of advantages. As mentioned previously, the IAT can be made accessible for non-human primates. It can be used for pre-schoolers, as they are not yet able to read. This is also the case for dyslectic or analphabetic people, including people with mental disabilities. Finally, a picture version could eliminate some of the issues inherent in translating the test when used with multiple cultures. According to Danziger and Ward (2010) the choice of language for an IAT can influence the results. In their study, they tested bilingual Arab-Israelis with an Arabic and Hebrew IAT and found that the Arabic version elicited a positive attitude effect for Arabs compared to Jews. The opposite was the case for the Jewish version. With a picture-only IAT the participants are no longer forced to think in the language used in the test. However, the cultural differences might still have an effect on the picture interpretation, as pictures are far from universal and their interpretation might depend on the cultural usage of pictures to the viewer (Jones & Hagen, 1980).

Word and picture differences

Despite its promise, a picture-only IAT would differ from IATs with words due to a number of factors. Although there seems to be no comparison of a picture-only IAT with an IAT with words, a handful of studies did investigate the effect of using pictures in another way. These studies compared IATs with pictures for the concepts (but not the attributes) with IATs with words for the concepts and attributes. Nosek, Banaji and Greenwald (2002) tested both a word-only race IAT with a word-picture race IAT. Although the word-only IAT score average was numerically higher than the word-picture IAT average, the authors did not test whether this difference was statistically significant. Chang and Mitchell (2011) did compare a word-only IAT and a picture-words IAT between subjects and found similar results, although this was not statistically significant. Dasgupta, McGhee, Greenwald and Banaji (2000) also compared a word-only IAT, but with a within-subjects design to test implicit white preference. They found that implicit white preference was significantly larger for the word-only version compared to the word and picture version ($p = 10^{-5}$). Foroni and Bel-Bahar (2010) compared a word-only IAT and a picture-word IAT twice: once in both their first experiment and second experiment. In both experiments the word-only IAT gave stronger results ($p < .01$ and $p = .05$ respectively).

Meissner and Rothermund (2015) noticed a trend in these few comparative studies and investigated for a potential cause. Although it seemed like there was a tendency for word-only IATs to provide stronger results, they hypothesized that this might not be due to the usage of words over pictures – the modality- itself, but the *match* between the usage of words or pictures for the concepts and attributes in IATs. The word-only IAT might be easier to process than a word-and-picture IAT. They compared two types of IATs: an insects/flowers IAT and an age-attitude IAT. For each IAT, they used either words for both concepts and

attitudes, pictures for concepts only, pictures for the attributes only, or pictures for both concepts and attitudes. In a series of experiments, they compared these IATs between two groups of 40 subjects. When comparing these IATs, they found evidence for the modality match hypothesis: the IAT effects were not exclusively higher for the word IAT. Instead, the effect was reversed when pictures were also used for the attributes. When the attributes were pictures, the IAT effect was the highest when pictures were used for the concepts, and lower when words were used for the concepts. Thus, exclusively using pictures for the IAT can provide results comparable with IATs that exclusively use words. In their analysis Meissner and Rothermund (2015) found that this is likely due to having to recode verbal and visual stimuli, which can mask the IAT effects. This recoding can simplify the task, which causes the IAT scores to increase. They also found a weaker effect with their second IAT, the age/attitude IAT. This was likely due to the difference in complexity of their concept and attribute picture stimuli. For their attribute stimuli they used pictures of complex scenes, whereas their concept pictures were clean high-contrast pictures of human faces. Additionally, the recoding strategy itself might change depending on the modality match, as both valence and salience can play a role in IATs depending on the stimuli used (Chang & Mitchell, 2011). This might also account for the weaker effect in their second IAT.

The present study

When taking the Meissner and Rothermund (2015) study into account, it stands to reason that picture-only IATs can potentially be as viable as word-only IATs. However, the possibility of factors that can diminish the IAT effect should be taken into consideration, such as the modality match of the picture or word stimuli and attributes. That is why a new picture-only IAT should be validated by comparing it to an IAT that uses words for the attributes. The Meissner and Rothermund (2015) study is a promising indication, but did not directly

compare the different IATs with the same subjects. As IAT scores can vary from person to person, it would be interesting to compare the IATs for individual subjects. Additionally, the groups for each IAT consisted of 40 participants. That is why the present study will therefore conduct a within-subjects comparison with both a picture-only IAT (P-IAT) and a word-picture IAT (W-IAT), using a large sample size. To compare the results of these implicit associations, this study will conduct an explicit race bias questionnaire between the two IATs. Although to my knowledge no study has compared a picture-only race P-IAT with a word-picture race W-IAT, a handful of studies did compare word-only IATs with word-picture IATs. Taking this previous research into account, it should be expected that the P-IAT provides IAT scores higher than the W-IAT, as the P-IAT has a modality match for the concepts and attributes (both pictures), whilst the W-IAT does not.

Methods

Participants

The study was conducted with 158 participants. After removing participants who did not finish the study, this was lowered to 141 (age M = 23.72, SD = 10.162, range 19-68). 27 participants were male (age M = 23.78, SD = 7.587, range 19-59), 114 female (age M = 23.70, SD = 10.709, range 19-68). All participants were 18 years or older, spoke Dutch as their native language and were of the Dutch nationality. Their parent(s) were also of Dutch nationality. 128 were right-handed (81.0%), 13 were left-handed (8.2%). All participants were recruited via the online recruitment system of Leiden University (Sona), flyers, posters and social media. The study has been approved by the local ethics committee on April 25th, 2018 #CEP18-0419/221.

Study Design

The study consisted of comparing the regular IAT that includes words (W-IAT) and a newer IAT where all stimuli consist of pictures (P-IAT). The study procedure followed a within-person design through an automated online task. The participants were screened for age, native language and the nationality of themselves and of both their parents. This is meant to ensure that they have all been exposed to standard Dutch culture and cultural biases and consider Caucasians their in-group and Moroccans their out-group.

After the introduction and screening, the participants started with one of either the W-IAT or P-IAT (counterbalanced), then filled in the race bias questionnaire, and finally completed the remaining IAT type. In either case the explicit race bias questionnaire (SRS) was conducted between the two IATs. In turn, each of these two groups used one of 4 possible W-IAT/P-IAT versions. This randomized whether the bias-compatible or bias-incompatible condition was done first, and whether the target on the right started as a positive or negative stimulus. These randomizations eliminated the possibility of an order effect for the two IAT tasks, for bias compatibility and for stimulus position.

Following the last IAT they were asked questions about their prior experience with people of Moroccan descent. With this data, a comparison between the P-IAT and W-IAT can be made whilst accounting for prior experience with the out-group and explicit race bias.

Procedure

Participants were given an access code to participate entirely online. The task was only available in Dutch. In order to spur participants to read the informational texts such as the introduction and debriefing, the option to continue was delayed with 5 seconds for smaller sections and 7 seconds for larger sections.

Participants were first given information about how their personal information is kept

confidential, of their right to stop participating, what the goal of the study entailed (comparing tasks) and how the study is designed. They were also provided with contact information to allow contact with the head of the COPAN research group or the researcher responsible for this study. After this the participants gave their consent for participating. Next, the participants were issued questions about their birth year, gender, handedness, native language, their own nationality, their father's nationality and their mother's nationality. The study would be terminated with a custom message if the participants indicated their age was lower than 18, if their native language was not Dutch, or if their own-, their father's- or their mother's nationality was not Dutch. This automated the selection process in order to create a select subgroup of participants.

After this, participants would receive instructions for the IAT task, followed by either the P-IAT or W-IAT (randomized). This was followed by the Symbolic Racism 2000 Scale (SRS), and lastly by the other IAT. After finishing these tasks, they were asked about their prior experience with people from Moroccan descent. If they answered that they knew one or more Dutch-Moroccans, they had to also indicate the extent to which they knew this person or these people on a scale. This scale ranged from 'I don't know this person/these people at all' (0) to 'I know this person/these people very well' (10).

When all questions and tasks were completed, the participants were debriefed. The debriefing made clear that the study would officially end after the debriefing was read. The debriefing informed the participants of what unconscious associations entail, what the greater goal of the study was (validation to enable comparative research) and how to reach the participants. Finally, the participants received the debriefing and were provided a link that enables them to provide the necessary information for receiving a study credit if they were first year psychology students. This was done in order to keep their responses anonymous.

15

Materials

Implicit Association Task   The Implicit Association Test or IAT (Greenwald, McGhee & Schwartz, 1998) is a task that measures the implicit associations for a concept with an attribute. The concept and attribute are both split in two categories. For this study a race bias IAT was used. In this IAT, for the concept of race (Caucasian and Moroccan categories) the strength of the implicit association with the attribute of valence (positive or negative) was measured. Thus, the IAT measured the reaction time for associating a race concept category with a valence attribute category. With the race IAT this can indicate a positivity or negativity bias towards the in-group (Caucasians) or the out-group (Moroccans). For example, if there is a negativity bias for an out-group, the association with the out-group concept and the negative attribute should be stronger and thus provide a faster reaction time.

This reaction time is measured during a categorization task that is carried out as fast as possible. For each trial the task provides a stimulus (bottom middle of screen) which must be categorized into either of two categories (top left and right). For this study the categorization is done by pressing the 'E' key for the left category and the 'I' key for the right category. See figure 1 for an illustration of the IAT procedure.
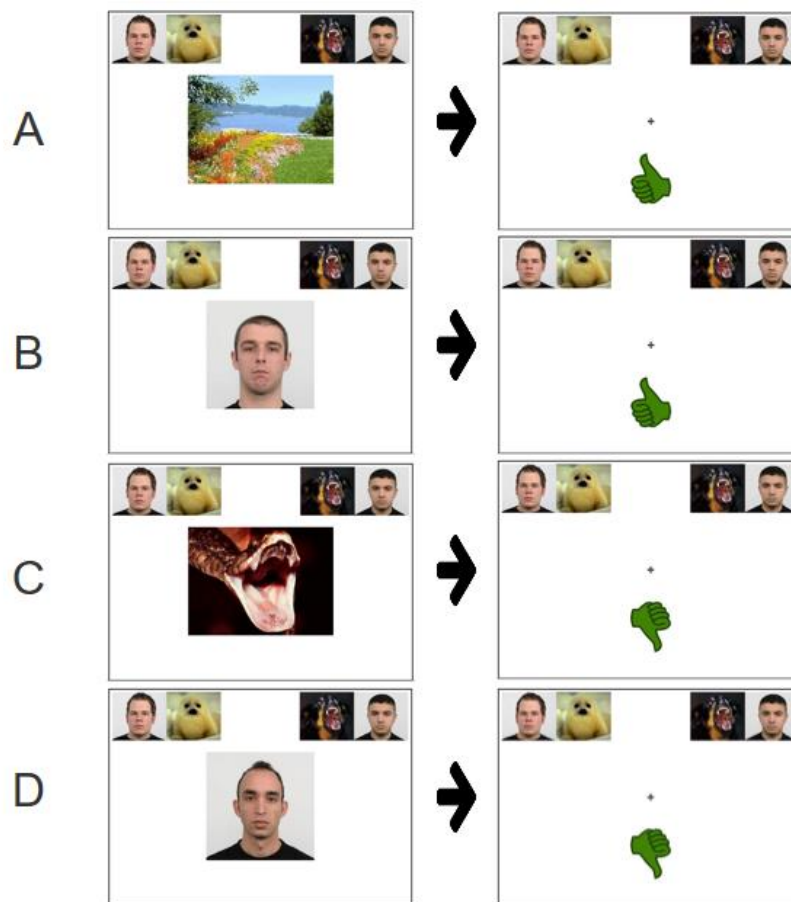
Figure 1. Scenario A to D illustrate the feedback for indicating

that the stimulus image belongs to the **left** category (done by

pressing 'E'). Shortly after the feedback the participant will be

presented with another stimulus image and they will have to

categorize yet again. This continues for 20 trials before the order

and combination of the categories above are changed.

In this study each IAT consisted of seven blocks. The first two blocks were training

blocks (40 trials total). The third and fourth block were experimental blocks (40 trials total).

The fifth block was a training block again (20 trials), and the final two blocks were

experimental blocks (40 trials total). The first block categorized the race concepts. The second

block categorized the valence attributes. The third block categorized both concepts and

attributes in combined categories. The fourth block did this as well, but switched the position of the attribute categories. The fifth block categorized valence attributes only. The sixth block categorized both concepts and attributes again, but had switched the concepts position from the third and fourth blocks. The seventh block did the same categorization, but had switched the attribute categories position from the sixth block. The randomization of this content is shown in the methods supplementation section (appendix A).

For both IAT tasks participants were issued one of four versions that varied on the following two randomized factors: the starting position of the concept and whether this is expected to be compatible or incompatible with in-group positivity bias. These four versions were: 1) compatible first (Caucasian concept on right with positive attribute), 2) incompatible first (Caucasian concept on right with negative attribute), 3) compatible first (Caucasian concept on left with positive attribute) and 4) incompatible first (Caucasian concept on left with negative attribute).

After completion, each IAT yielded separate reaction times. For both versions of the IAT the reaction times were removed that exceeded 10.000 ms or under 300 ms for more than 10% of trials, in accordance to Greenwald, Banaji and Nosek (2003). With these reaction times, a difference score (D-score) was calculated for each IAT and for each participant. D-scores of two standard deviations above or below the mean were deemed to be outliers.

The current study included two versions of the IAT: the P-IAT and the W-IAT. Both versions functioned in the same way except for their key differences in the usage of words and pictures. In the next section these two versions will be outlined individually.

W-IAT The word IAT or W-IAT is a type of Implicit Association Test (Greenwald et al., 1998) that is used commonly. In the current study the term W-IAT is used to refer to the IAT that uses words for the categories and pictures for the stimuli. For the categories, the

concepts (Caucasian and Moroccan) were written in a black font as Dutch ('Nederlands') or Moroccan ('Marokkaans'). The attributes were written in a green font as positive ('positief') or negative ('negatief'). In combined category trials the black concept words and green concept words were written on top of each other with 'or' ('of') between them. Their order (top/bottom) was randomized. For the stimuli, the W-IAT used the same images as the P-IAT, which were taken from the Radbout Facial Database (Langner et al., 2010). For consistency the pictures that are used for the categories in the P-IAT are also absent amongst the stimuli used in the W-IAT.

P-IAT   The P-IAT is an IAT (Greenwald et al., 1998) that uses pictures for both the categories and stimuli. For the concepts the P-IAT used a subset from the Radboud Faces Database (RaFD, Langner et al., 2010) that consists of seven Moroccan and seven Caucasian faces. These all sport a neutral expression and gaze at the camera. The pictures have been validated for their important characteristics of the intensity, clarity, genuineness, expressiveness, attractiveness and valence of expression. Additionally, the pictures have been controlled for facial expression, gaze direction, head orientation, representation of gender, representation of both adults and children, lighting conditions, the position of facial landmarks, and image background. For the attributes the P-IAT uses pictures from the International Affective Picture System (IAPS, Lang, Bradley & Cuthbert, 2008). Six of these depict a positive and six a negative object, animal, or natural scene. Excluding pictures that include human faces, these were respectively the most positively and most negatively rated pictures from the database.

For the category indicator pictures the P-IAT used one picture from each category. This picture was only used here and did not appear amongst the stimuli. In combined category trials each dual category was represented by a concept and attribute category picture, which

were placed next to each other. Their position (inside/outside) was randomized. For the stimuli the P-IAT used the six remaining Moroccan face pictures, six remaining Caucasian face pictures, five remaining positive affect pictures and five remaining negative affect pictures. The picture stimuli used for the P-IAT are included in Appendix B.

Symbolic Racism 2000 Scale (SRS)

The Symbolic Racism 2000 Scale (Henry & Sears, 2002) measures explicit race bias. The study used a Dutch translation, which is not yet validated. In this version 'blacks' had been switched with 'Marokkanen' (Moroccans) and 'whites' had been switched with 'Nederlanders' (the Dutch). Furthermore, the sociodemographic groups the Irish, Italians and Jews from the USA had been translated to the Surinamese and Polish groups living in the Netherlands.

The SRS consists of 8 statements with 4 possible opinions for each statement. Participants are asked to select the opinion that would be the closest to their own opinion. These are on a 4-point scale with the answers differing for each question. In this translated version each statement is about Moroccans in the Netherlands. A transcript of the translated questionnaire has been included in appendix C (in Dutch).

Results

Out of the 158 initial participants, 17 participants were excluded due to not finishing the study. Due to these exclusions, the analysis was conducted with the remaining 141 participants. For a further 20 participants individual P-IAT and/or W-IAT D-scores were automatically excluded due to not meeting the reaction time and error minima. This was the case when their reaction times exceeded 10.000 ms, or were shorter than 300 ms for more than 10% of their trials. These criteria were based on Greenwald, Banaji and Nosek (2003).

Within this sample 2 W-IAT D-score outliers (.016 %) and 1 P-IAT D-score outlier (.008 %) were removed for exceeding the reaction-time filter criteria based on the median absolute deviation (MAD) method (Leys, Ley, Klein, Bernard & Licata, 2013). The current study is moderately conservative (-2.5 MAD $< x_i <$ 2.5 MAD) to balance out the improvement of the data over the overall loss of data. The W-IAT D-scores were -0.66 (-2.65 MAD from median of .5763) and -0.71 (-2.76 MAD from median of .5763), and the P-IAT D-scores were -0.58 (-2.56 MAD from median of .4174). Additionally, this decision was based on comparisons of the improvements in the histogram and Q-Q graphs for each possible level of conservatism. These histograms and Q-Q plots can be viewed in Appendix D.

To test whether the W-IAT and P-IAT D-scores differed significantly from 0, two one-sample t-tests were conducted. The D-scores were normally distributed, as assessed by the Shapiro-Wilk's test for both the W-IAT D-scores ($p = .492$) and P-IAT D-scores ($p = .185$). The distributions are shown in figure 2 and 3 for the W-IAT and P-IAT respectively. Due to exceeding the moderately conservative criteria, two outliers were removed from the W-IAT D-scores, and one was removed from the P-IAT D-scores. In the ensuing analysis, there were no outliers, based on inspection of a boxplot and Q-Q plot. Firstly, a one-sample t-test was conducted with the W-IAT. The mean D-score ($M = .53$, $SD = .40$) was higher than 0, a statistically significant mean difference of .53, 95% CI [.478 to .618], $t(128) = 15.4$, $p < .001$. Thus, the participants were significantly faster when categorizing a congruent combination of concepts and attributes compared to an incongruent combination when using an IAT with words for attributes. Secondly, a one-sample t-test was conducted with the P-IAT. The mean D-score ($M = .41$, $SD = .36$) was higher than 0, a statistically significant mean difference of .41, 95% CI [.343 to .471], $t(124) = 12.7$, $p < .001$. Thus, the participants were significantly faster when categorizing a congruent combination of concept and attributes compared to an incongruent combination when using a picture IAT.
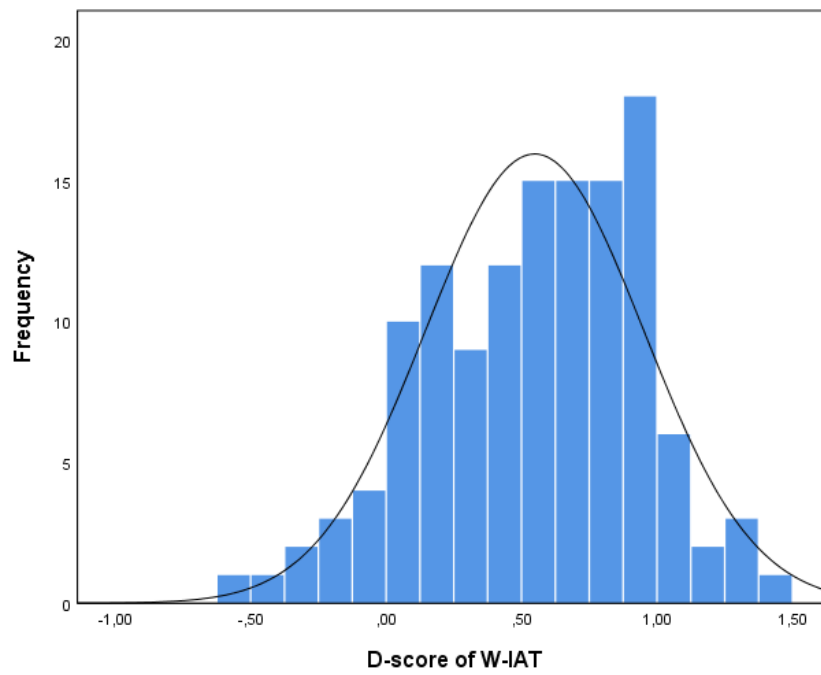
Figure 2. The frequency distribution of W-IAT D-scores
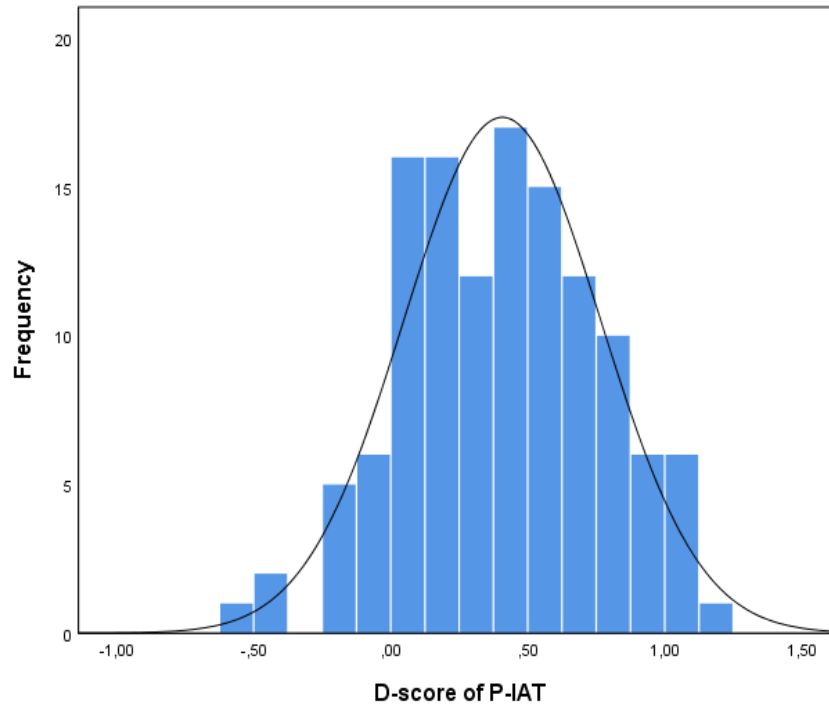
Figure 3. The frequency distribution of P-IAT D-scores

Next, a paired samples t-test was conducted with both the P-IAT and W-IAT D-scores. The mean P-IAT D-score (M = .41, SD = .35) was lower than the mean W-IAT D-score (M = .53, SD = .40), a statistically significant mean difference of -.12, 95% CI [-.202 to -.044], $t(118)$ = -3.092, p < .05. Thus, the implicit association bias was reportedly stronger when measured with the W-IAT compared to when measured with the P-IAT.

To test for a relationship between the two versions, a Pearson correlation coefficient was computed to assess the relationship between the P-IAT D-scores and the W-IAT D-scores. A significant positive correlation between the two variables was found, r = .344, N = 125, p < .001. Overall, there was a medium (Cohen, 1992) positive relationship between implicit bias scores measured by the P-IAT and as measured by the W-IAT. A scatterplot summarizes the results in figure 4.
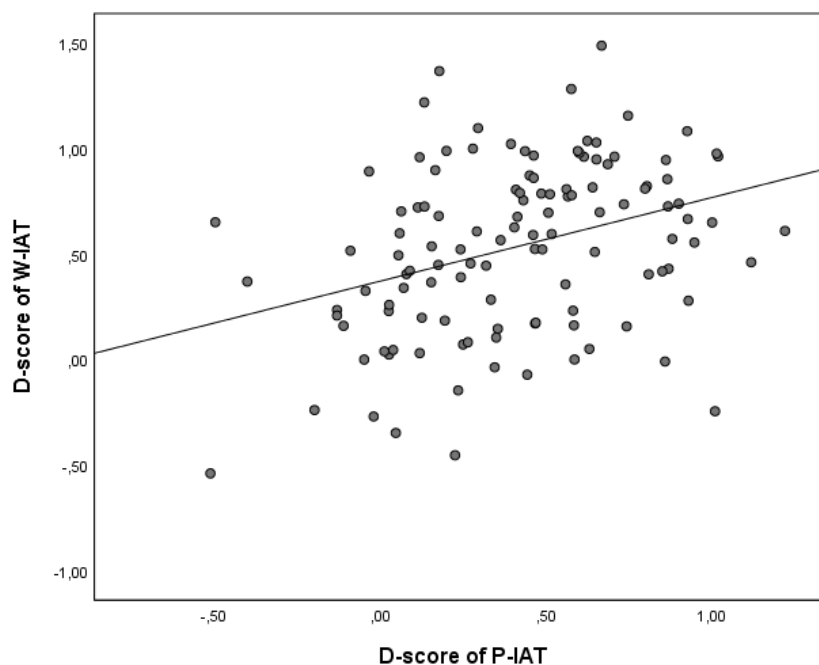

Figure 4. Scatterplot of W-IAT D-scores and P-IAT D-scores.

In addition to comparing the IATs with each other, the IATs were also compared with the explicit bias measure, the SRS. In the data analysis, the SRS was converted to a continuous 0 to 1 scale in accordance with Henry & Sears (2002). A Pearson correlation coefficient was computed to assess the relationship between explicit bias (as measured by the SRS) and implicit bias as measured by the W-IAT and P-IAT. Firstly, a significant positive correlation was found between the SRS scores and W-IAT D-scores, $r = .277$, $N = 129$, $p < .005$. Overall, there was a medium positive relationship between implicit bias scores measured by the W-IAT and explicit bias scores. Increases in explicit bias were correlated with increases in implicit bias as measured by the W-IAT. A scatterplot summarizes the results in figure 5. Secondly, a non-significant correlation was found between the SRS scores and P-IAT D-scores, $r = -.001$, $N = 125$, $p = .990$. Overall, no relationship was found between implicit bias scores measured by the P-IAT, and explicit bias scores. A scatterplot summarizes the results in figure 6.
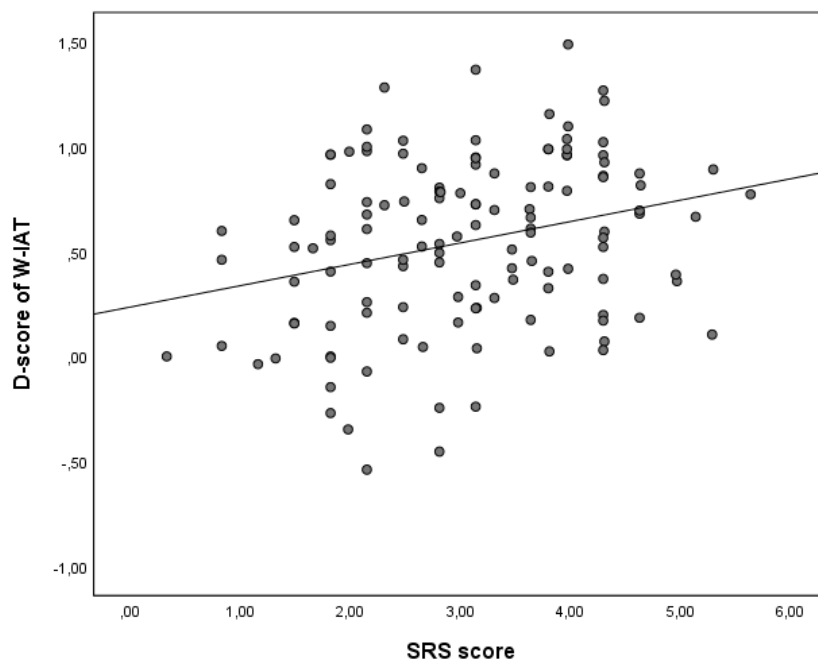

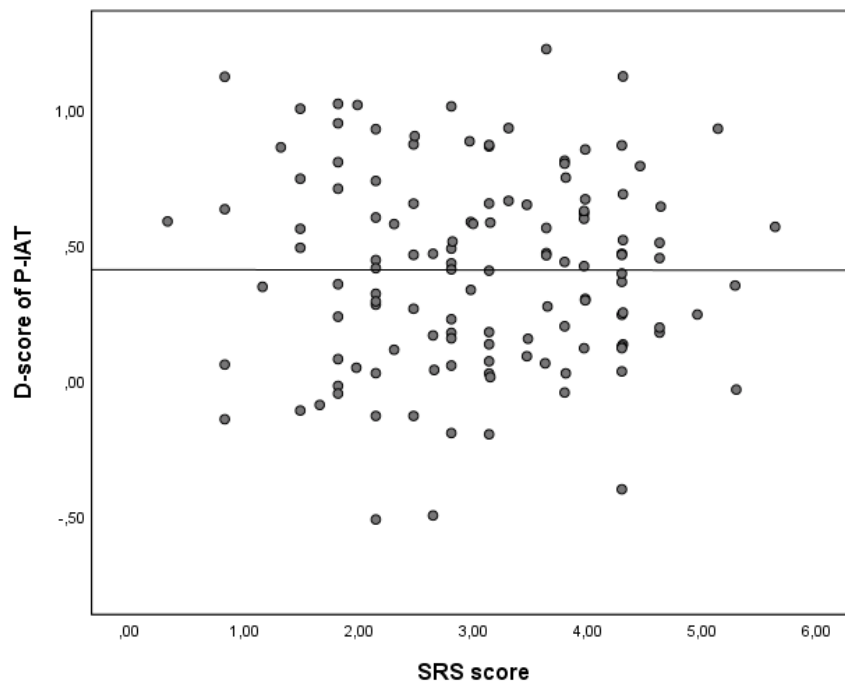Figure 5. Scatterplot of the SRS scores and W-IAT D-scores.

Figure 6. Scatterplot of the SRS scores and P-IAT D-scores.

The final variable to be tested for was the effect of outgroup familiarity on the IATs, which was enquired in the online questionnaire. This variable is known to influence implicit and explicit biases (Pettigrew & Tropp, 2006) and thus might underlie the IAT and SRS results. A nonparametric test was used due to the non-normality caused by the usage of the option 0, with which participants could indicate that they knew no people from the outgroup whatsoever. A Spearman's rank-order correlation coefficient was computed for the effect of reported outgroup familiarity on W-IAT D-scores, P-IAT D-scores, and SRS scores. First, a non-significant negative correlation was found with the W-IAT D-scores, $r = -.129$, $N = 129$, $p = .091$. Thus, there was no significant relationship with implicit bias as measured by the W-IAT. Second, a non-significant negative correlation was found with the P-IAT D-scores, $r_s = -.130$, $N = 125$, $p = .147$. Thus, there was no significant relationship with implicit bias as measured by the P-IAT. Third, a significant negative correlation was found with the SRS scores, $r_s = -.173$, $N = 141$, $p < .05$. Thus, there was a negative, small (Cohen, 1992) yet

25

significant relationship between reported outgroup familiarity and explicit bias as measured

with the SRS. The scatterplots of the relationship of reported outgroup familiarity with W-

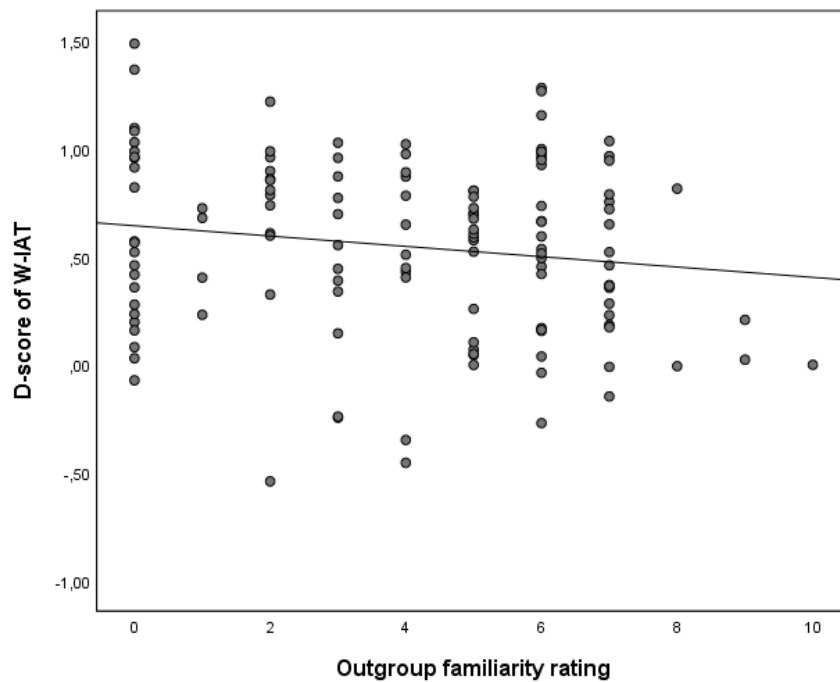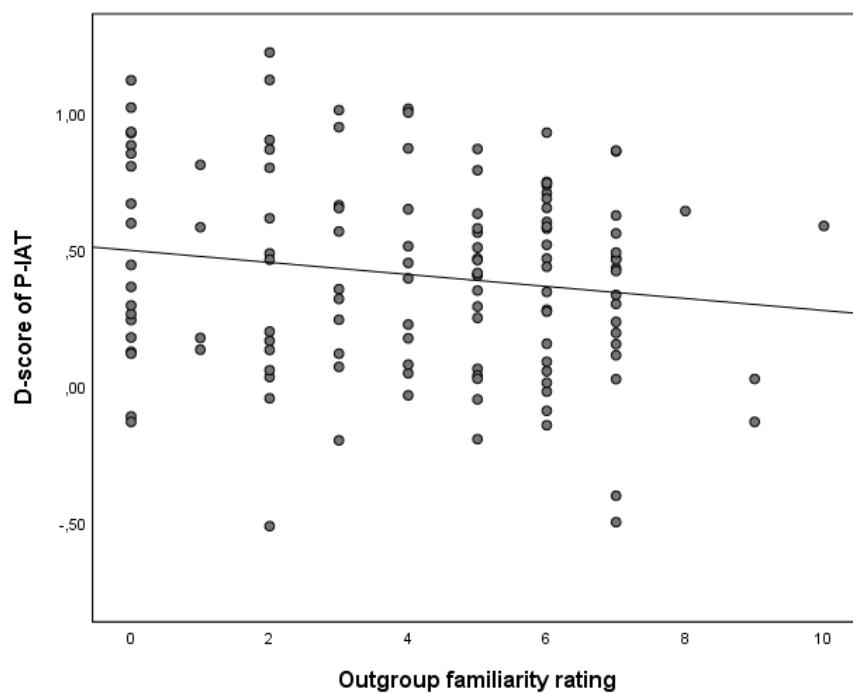IAT, P-IAT and SRS scores are shown in figure 7, 8 and 9 respectively.

Figure 7. Scatterplot of outgroup familiarity ratings and W-IAT D-scores.

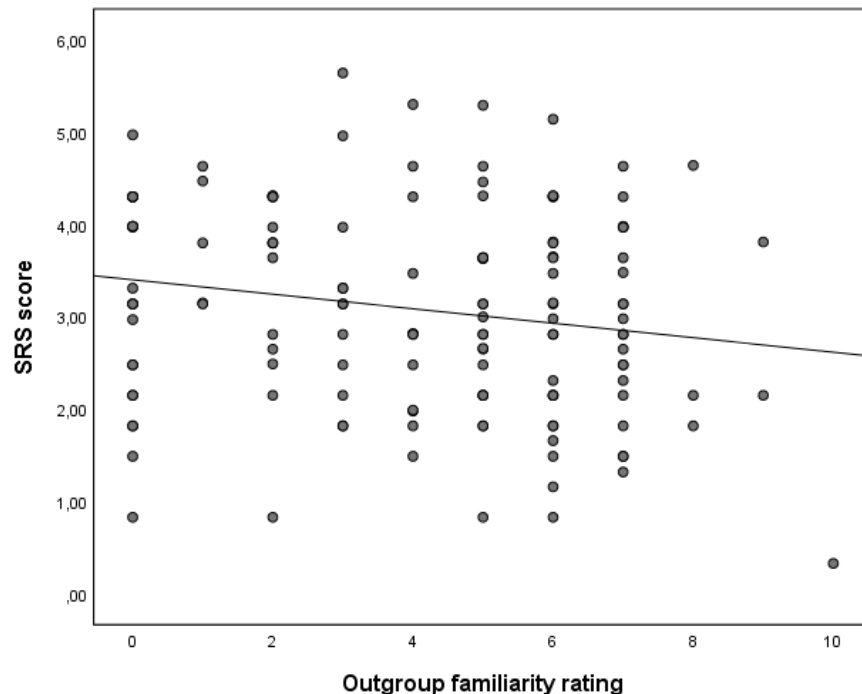Figure 8. Scatterplot of outgroup familiarity ratings and P-IAT D-scores.

Figure 9. Scatterplot of outgroup familiarity ratings and SRS scores.

After the tests were concluded, three more post-hoc tests were conducted to look into the difference between the relationship of the W-IAT and P-IAT with the SRS measure. One possible reason for this discrepancy might be that the groups differ due to the removal of participants that did not finish the study completely. As these participants were originally included in the randomization procedure, this might have resulted in uneven groups. In order to test this, three independent samples t-tests were conducted for the W-IAT, P-IAT and SRS measures. For each measure the subset that started with the W-IAT (W-IAT first group, $N = 72$) was compared with the subset that started with the P-IAT (P-IAT first group, $N = 69$). The D-score distributions and SRS score distribution were approximately normally distributed.

Firstly, the W-IAT D-scores were not significantly higher for the W-IAT first subset

(M = .558, SD = .413) than for the P-IAT first subset (M = .538, SD = .395), $t(127)$ = -.283, p = .778. Levene's Test did not indicate unequal variances (F = .397, p = .530). Secondly, the P-IAT D-scores were not significantly higher for the P-IAT first subset (M = .432, SD = .416) than for the W-IAT first subset (M = .384, SD = .300), $t(107)$ = .738, p = .462. Levene's Test indicated unequal variances (F = 6.465, p = .012), so degrees of freedom were adjusted from 123 to 107. Thirdly, the SRS scores were not significantly higher for the W-IAT first subset (M = 3.211, SD = 1.092) than for the P-IAT first subset (M = 2.944, SD = 1.097), $t(139)$ = -1.447, p = .150. Levene's Test did not indicate unequal variances (F = .125, p = .724). The histograms of the two P-IAT subsets, the two W-IAT subsets and the two SRS subsets can be viewed in figure 10, 11 and 12 respectively.
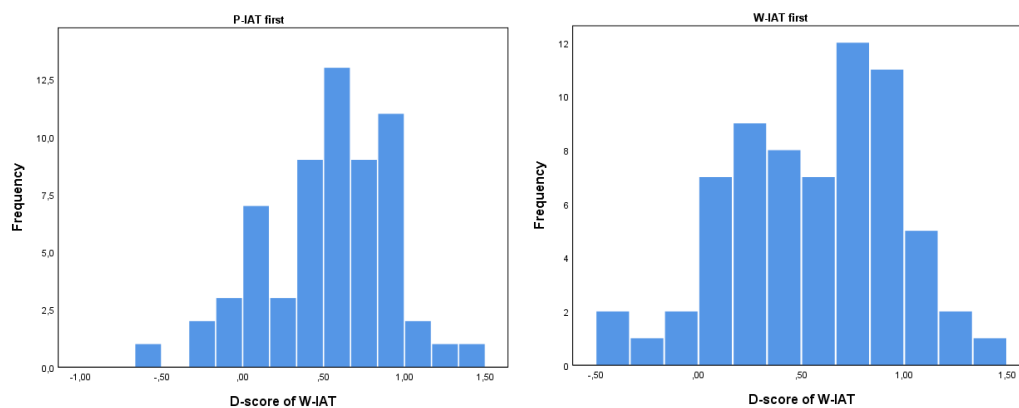

Figure 10. Histograms of W-IAT D-Scores of the P-IAT first (left) and W-IAT first (right) subgroups.
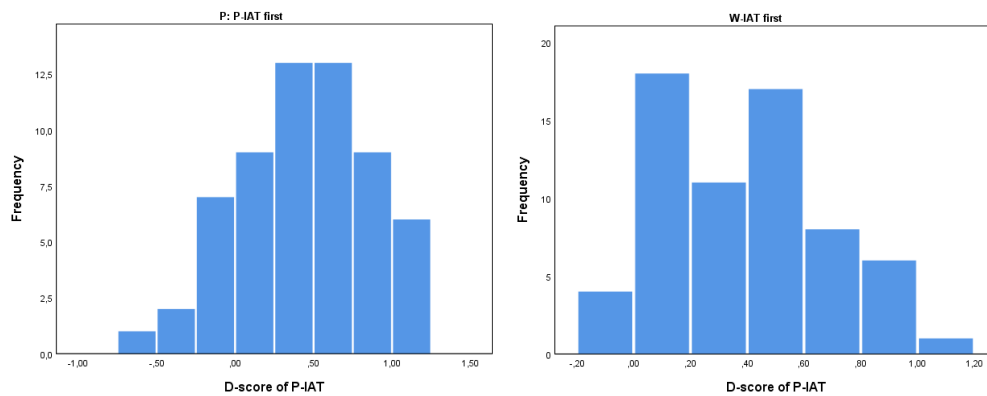
Figure 11. Histograms of P-IAT D-Scores of the P-IAT first (left) and W-IAT first (right) subgroups.
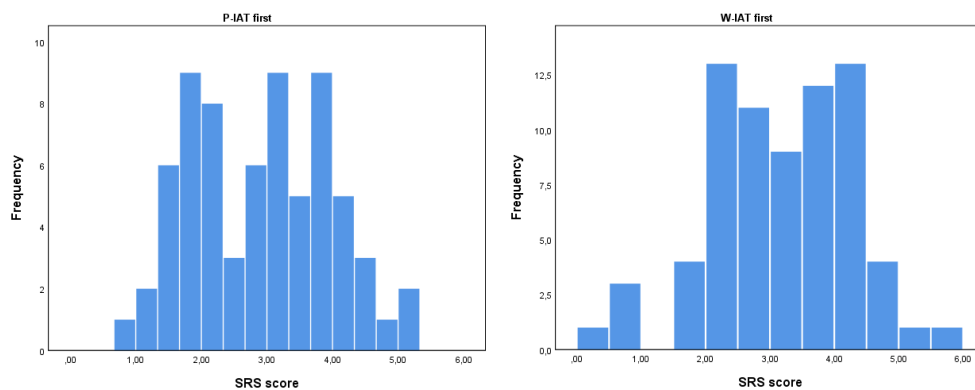
Figure 12. Histograms of SRS scores of the P-IAT first (left) and W-IAT first (right) subgroups.

Discussion

The aim of this study was to conduct a within-person comparison of a picture-only IAT with an IAT that uses words for the attributes. Overall, a significant race bias was found with both the W-IAT and the P-IAT. Contrary to expectations, the W-IAT indicated a stronger race bias than the P-IAT did. Moreover, the SRS was significantly correlated with the W-IAT, but not with the P-IAT. Finally, the reported outgroup familiarity had a significant negative correlation with the SRS, but was not significantly correlated with the W-

29

IAT or P-IAT. Although both versions of the IAT indicated the presence of race bias, they did not do so equally. The P-IAT indicated a smaller average race bias than the W-IAT, contrary to our hypothesis. This could be the case for a variety of reasons.

P-IAT stimuli

One reason could be the content of the P-IAT itself. The W-IAT has been improved over the years, whereas the P-IAT is currently still a rare task. The lack of refinement in the execution of the P-IAT might explain the different results for the task. Besides the inherent differences between words and pictures, the subtle differences due to their content could also be a factor. A particular way this could have occurred is that in the P-IAT task, the stimuli were drawn from the pool of pictures used for each category. However, their representativeness has not been tested separately. Unlike the pictures they could not be randomized, as each category was represented with one picture each. Thus, individually the two pictures of faces used might not have been representative enough for their respective groups. To my knowledge the RaFD (Langner et al., 2010) pictures used for the P-IAT have not been tested on representativeness for their given group. The faces themselves have been validated on various important facial characteristics (Langner et al., 2010) and on perceived big five traits (Jaeger, 2018). However, the groups of Caucasian and Moroccan face pictures have not been directly compared, to my knowledge. However, the validation data of Langner et al. (2010) is freely available for all images. This data includes the agreement percentage, expression intensity, expression clarity, expression genuineness and valence rating for each image. In a future study, it would be recommended to control the choice of P-IAT stimuli on these factors to ensure the pictures used are comparable within and between each race category.

Furthermore, the face picture stimuli differed in a number of ways from the affective

picture stimuli. The face picture stimuli were all purposefully shot in the same way to remove the influence of important factors such as lighting, angle, expression, clothing etc. On the other hand, the affective stimuli depicted various objects, scenes and animals in different ways. Although their purpose was to elicit emotions, their presentation was not homogenized like the face pictures. This discrepancy could have influenced in distinct ways how these two types of pictures were recoded into the abstract concepts they depict. This could have reduced the advantage of a congruent stimulus modality that the P-IAT was hypothesized to have compared to the incongruent stimulus modality of the W-IAT. Systematic differences in the salience of pictures on visual aspects such as the lighting, colour and composition cannot be ruled out. These systematic differences could have possibly allowed recoding.

Another important factor that relates to the P-IAT stimuli choice is that the race attribute was represented with faces whereas the valence attribute was represented with pictures of animals, scenes and objects. Human faces are encoded by the specialized Fusiform Face Area whereas non-face stimuli are not. This is a distinct process from scene encoding, which is similarly encoded by a functionally specialized area: the para-hippocampal place area (Kanwisher, 2010). The effect this difference might exert on the W-IAT and P-IAT results might be cause for future research. Furthermore, in contrast to the P-IAT the W-IAT used words to represent both race and valence. In order to eliminate this discrepancy, further studies could choose to use positive and negative expressions to represent the attribute of valence. The RaFD (Langner et al., 2010) includes pictures that portray a variety of emotions, including joy and anger and might be useful for this purpose.

Modality match

A second reason for the difference in performance of the two IATs could lie in the differences between the current study and the study by Meissner and Rothermund (2015), as

they also compared picture-only IATs with word-picture (and word-only) IATs. They had participants conduct one of the four combinations of using words or pictures for the concepts and attributes for a flower/insect or old/young IAT. The flower/insect task provided higher scores for the picture-only version compared to the version with words for attributes. The old/young IAT provided a slight but non-significant difference between these two versions. However, in the current study the opposite was found. This could be for a number of reasons.

One reason is the usage of different IATs. Meissner and Rothermund (2015) noted that they chose to use the old/young version to represent IATs that measure implicit attitudes to social groups. The difference between their non-social and social domain IAT could indicate that the usage of social stimuli itself could influence the recoding process. This relates to a potential influence on the P-IAT/W-IAT difference mentioned in the section above: the race attribute stimuli used faces (social), whereas the valence attribute stimuli did not use faces (non-social). This might be due to recoding (de Houwer, 2003). Instead of categorizing based on the desired attribute, participants might focus on another attribute present for a category. There are no salient differences in the presentation of valence attribute words and race attribute words, whereas valence attribute pictures are non-social stimuli and race attribute pictures are social stimuli. If the usage of social stimuli influences the social process differently than non-social stimuli, this might have differentially influenced the recoding and therefore performance on the W-IAT and PIAT.

However, caution should be taken when comparing the current study with the Meissner and Rothermund (2015) study, as there are two more differences in the study design of the studies. Besides the IAT choice, a second discrepancy with their study is that they used a between-subjects design, whereas the current study is a within-subjects design. Compared to this between-subjects design, the recoding process in the current study could possibly have been influenced by learning effects. Despite the randomization of the P-IAT and W-IAT tasks,

they were still conducted in successive order, which was not the case in Meissner and Rothermund's (2015) between-subjects study. This might make it more difficult to directly compare with their results.

A third discrepancy between the studies is that the Meissner and Rothermund (2015) study makes use of the ReAL model (Meissner & Rothermund, 2013), which is a multinomial model that maps the different ways information is processed in a way that can distinguish between dissociative associations and factors influencing recoding. Due to the ReAL model, Meissner and Rothermund (2015) modified their IAT procedures and data analysis in accordance with Meissner and Rothermund (2013). These modifications required participants to respond more quickly and feature a response deadline which was continuously updated based on the errors made in the previous block pair. When the deadline is not met, a red rectangle framed the stimuli to remind participants that they should respond more quickly. These modifications were designed to increase the error rate to yield the amount of errors needed to improve the reliable estimation of the ReAL model parameters. Meissner and Rothermund (2015) also conducted a replication study for the flower/insect IAT that used the non-modified procedure and computed their scores based on the D-score algorithm from Greenwald et al. (2003), which showed the same modality effect. However, their young/old IAT was not replicated this way, and when they conducted a version with a fixed response deadline the modality effect vanished when pictures were used as attributes.

These methodological differences make it harder to directly compare the Meissner and Rothermund (2015) study with the current implicit race bias IAT study, as their only social bias IAT makes use of modifications for the ReAL model analysis and is conducted between subjects. Regardless, these differences might partially explain the differing results.

Levels of representation

A third reason the W-IAT and P-IAT might have differed is that the current study did not account for the level of representation of the picture stimuli compared to word stimuli. Category label stimuli can differ on the extent to which they represent the categories. This level of representation (LR) can invoke a varying range of exemplars, which can influence the perception of ingroups and outgroups (Park, Ryan & Judd, 1992). Foroni and Bel-Bahar (2010) applied this concept of levels of representation to implicit bias in comparing IATs with words only and IATs with pictures for the concepts and words for the categories (which they call PIATs, but is called a W-IAT in the current study). A label such as 'Dutch' seems to be more inherently representative of a group than a picture of the face of any individual. This would mean words are on a higher level of representation (LR) of stimuli than pictures. Foroni and Bel-Bahar (2010) tested their hypothesis with two experiments. In their first experiment, they compared two race IATs which are identical save for the usage of pictures or words for the category stimuli. They found that the word-only IAT yielded higher scores than the IAT that used pictures for the categories. In the second experiment, they manipulated the difference of stimulus LR for different sets of stimuli for a race IAT. The results indicated that stimulus type can influence the height of the IAT score. Moreover, the results suggested that stimulus LR was more relevant to scores than stimulus modality, although the usage of different pre-existing tests could have influenced this effect, as the tests were not pre-tested on stimulus LR. Therefore, in the third experiment Foroni and Bel-Bahar (2010) tested pre-tested material on one modality (written text) to compare basic-level and subordinate-level stimuli. Again, they found that stimulus LR was an influence despite modality. In their fourth experiment, they also found this with social categories comparing two IATs with pictures for stimuli and words for the categories.

In the current study, stimulus LR has not been taken into account despite being a possible factor to consider. The W-IAT used words such as 'positief' (positive) and 'negatief' (negative), whilst the P-IAT used pictures of specific instances that are generally deemed positive or negative, such as pictures of a baby seal or a dog growling threateningly. These pictures might be less representative of 'positive' and 'negative', and thus have a lower stimulus LR than the words themselves.

Although stimulus LR seems to work even when stimulus modality is kept constant, it might be difficult to remove the effect due to the inherent specificity of information pictures have when depicting a person. However, this was taken into account in the fourth experiment in the Foroni and Bel-Bahar (2010) study. In order to manipulate the LR between two otherwise identical IATs with pictures for stimuli, they compared pictures with individuals with pictures of groups of individuals. This raises the stimulus LR, as the pictures now represent groups instead of specific instances from a group. For future research it would be interesting to compare picture-only IATs that use groups for the stimuli to IATs that use words in further studies.

Explicit bias differences

Besides the previously mentioned differences, another apparent difference between the two IAT versions is their correlation with the SRS, which was correlated with the W-IAT, but not the P-IAT. One reason for this discrepancy might lie at the surface level of the measures: the SRS uses written words. As mentioned previously, words are on a higher level of stimulus representation than pictures (Foroni & Bel-Bahar, 2010), as they can represent broader concepts such as groups compared to pictures of specific individuals. Similar to the W-IAT, the SRS uses word labels to denote social groups due to being a written questionnaire. This might be a possible explanation for the discrepancy between the W-IAT and P-IAT.

However, this might not be the only reason, as explicit bias reports have a complex relationship with implicit bias test results. Although differences in social sensitivity elicited by various IATs can explain some of this discrepancy in meta-analyses of the explicit-implicit relation literature (Greenwald, Poehlman, Uhlman & Banaji, 2009), this is far from the only influence on the implicit-explicit relationship. Implicit and explicit associations are two distinct yet still related processes. According to the the MODE model (Fazio, 1990), the motivation and opportunity to change explicit attitudes can increase the difference between reported implicit and explicit bias. Hofmann et al. (2005a) proposed a model that expanded on this notion by including other dual process theories, positing that implicit and explicit processes are distinct concepts that can influence each other and their assessments in various ways. This model (see figure 13) consists of four key parts: explicit (mental) representation, implicit (mental) representation, the explicit indicator (e.g. an explicit bias score) and the implicit indicator (e.g. an IAT D-score). This model distinguishes five points at which moderators can influence these four constructs: 1) the translation between explicit representation and implicit representation; 2) the additional information that can change the explicit representation; 3) explicit assessment factors; 4) implicit assessment factors; and 5) study design factors.
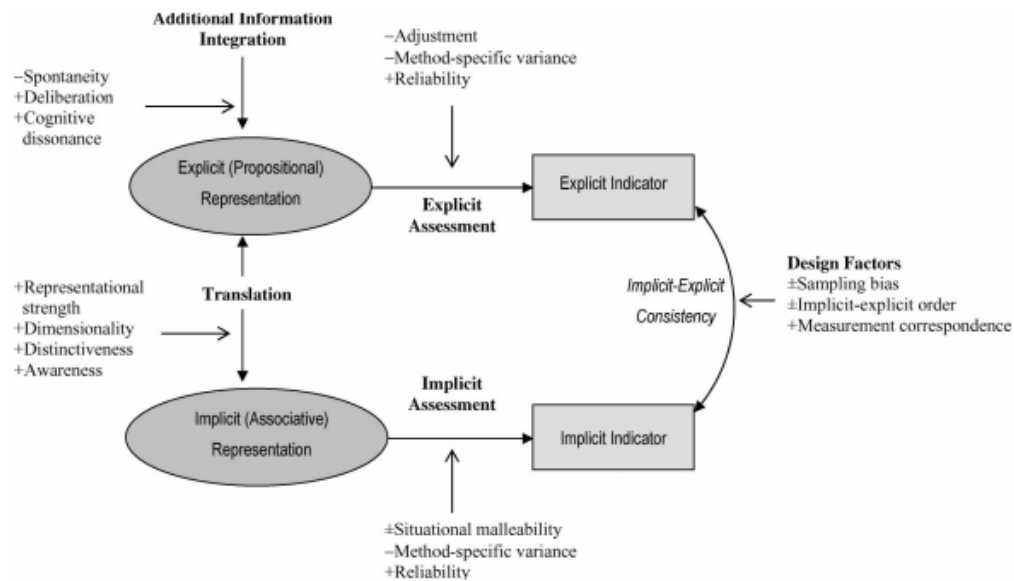
Figure 13. The implicit-explicit consistency model by Hofmann et al. (2005a, p. 343).

The first group of moderators lie between the implicit and explicit representations. This relation involves mental representations and thus takes place within subjects. As the current study used a randomized within-subjects comparison for the implicit and explicit measures, these moderators are unlikely to underlie the difference in correlations of the W-IAT and P-IAT with the SRS. However, one moderator of note mentioned by Hofmann et al. (2005a) and Nosek (2005) is whether the dimensionality of the evaluation is bipolar or unipolar. In bipolar evaluations one attitude is directly opposed to the other, and being 'for' it would mean being 'against' the other attitude. In unipolar evaluations this is not the case. According to Nosek (2005), this more complicated structure might make retrieving mental representations more difficult, reducing the correlation between explicit and implicit bias. A negative outgroup race bias does not necessarily create a more positive ingroup race bias. Thus, race bias is a unipolar evaluation to an extent, which would partially explain the implicit-explicit difference.

The second group of moderators involve the integration of extra information for explicit mental representations. The lack of spontaneity and the opportunity for deliberation during the explicit measure would give the respondents the opportunity and motivation to change their explicit mental representation (Fazio, 1990). This means that implicit-explicit differences can only arise if there is an opportunity and motivation to alter the explicit indicator's result. Another moderator is cognitive dissonance. If participants experience cognitive dissonance due to their actions during the explicit measure, they would be motivated to resolve this by changing their explicit attitude and rely less on their implicit associations (Gawronski & Strack, 2004). As the current study was conducted online without supervision, it did not account for spontaneity, deliberation and cognitive dissonance during the SRS. The occurrence of these moderators during the SRS might further explain the discrepancy between the correlation of the SRS with the W-IAT and P-IAT.

The third group of moderators involves the direct measurement of explicit attitudes. Hofmann et al. (2005a) mention that adjustment is a major moderator for this domain, as it is during measurement that responses can be adjusted for explicit measures, but not implicit measures. Again, this might be a possible contributor to the discrepancy between SRS and IAT correlations. Another possible moderator is the reliability of the measure itself. The SRS is internally consistent (Henry & Sears, 2000), and thus has sufficient reliability. Therefore, it seems unlikely that unreliable measurement was a contributor.

The fourth group of moderators involve the direct measurement of implicit attitudes. According to Hofmann et al. (2005a), the situational malleability, method-specific variance and reliability of the implicit assessment can further moderate the relation between implicit and explicit attitude scores. The situational malleability refers to the pre-activation of associations by priming before the measurement, which can change the implicit association during a test (Wittenbrink, Judd & Park, 2001). As the current study was conducted online,

this could have influenced the implicit-explicit correlations. The other two moderators - method-specific variance and reliability- are more dependent on the specific version of the measures used. In the current study, this would relate to the notion that the W-IAT has been tested more rigorously and has been perfected for longer than the P-IAT.

The fifth and final group of moderators involves the study design factors. Hofmann et al. (2005a) names sampling bias, implicit-explicit order and measurement correspondence as known implicit-explicit relation moderators in this domain. The occurrence of a sampling bias between two groups for implicit and explicit data has been prevented by using a within-subjects design. The implicit-explicit measurement order has been nullified by randomization, as there is no significant difference between the SRS of the W-IAT-first and P-IAT-first groups. However, the measurement correspondence of the explicit and implicit measures might be of influence, as this is known as a predictor of attitude and behaviour consistency (Hofmann, Gawronski, Gschwendner, Le & Schmitt, 2005b). As mentioned previously, a possible explanation can be found in that both the W-IAT and SRS involve verbal processing, whereas the P-IAT specifically lacks verbal processing.

Using this framework, various possible moderators come to light. However, a post-hoc analysis that investigated the order-effects of the W-IAT and P-IAT revealed that the SRS performance was not significantly different for the W-IAT-first and P-IAT-first groups. This evidence seems to discredit the possibility of a direct influence of either IAT on the SRS scores, as this indicates that the randomizations was successful. Although there were non-significant differences in SRS scores between the P-IAT first and W-IAT first groups, this could be due to a standard gradual decrease in reaction time due to mental fatigue (Möckel, Beste & Wascher, 2015) between the two versions. Despite these findings, the previously mentioned moderators would explain why the W-IAT, but not the P-IAT was correlated with the SRS results.

Outgroup familiarity

The W-IAT and P-IAT had no significant relationship with familiarity with the outgroup. Outgroup familiarity did have a relationship with the SRS, although this was weak at best. This relation is not entirely in line with the literature. In their influential meta-analytical study, Pettigrew and Tropp (2006) found that outgroup contact typically reduces prejudice. Contact was found to reduce both implicit and explicit attitudes (Aberson, & Haag, 2007) and modulate racial bias throughout the lifespan (Kubota, Peiso, Marcum & Cloutier, 2017). Intergroup friendship (Aberson, Shoemaker & Tomolillo, 2010; Turner, Hewstond & Voci, 2007) and living together (Burns, Corno & La Ferrara, 2015) have also been found to reduce implicit and explicit bias. However, a recent meta-analytical study (Paluck, Green & Green, 2018) notes that the effect of contact on prejudice can vary due to a variety of factors. It should be noted that in the two aforementioned meta-analytical studies contact was defined as "actual face-to-face interaction between members of clearly defined groups" (Pettigrew & Tropp, 2006, p. 754; Paluck Green and Green, 2018, p. 8). In the current study participants were asked to rate the extent to which they know one person or multiple people from the outgroup. It is not clear what their relationships entail. Thus, the exact mechanisms in which their reported outgroup familiarity is related to their performance might vary beyond actual face-to-face *contact*.

Interestingly, Aberson and Haag (2007) found evidence that indicates that implicit associations were affected by outgroup contact independently from explicit associations. They found several mediation relationships between explicit attitudes and stereotyping, whereas they found no mediators for contact and implicit attitudes. This was similar to the findings by Tam, Hewstone, Harwood and Voci (2006), who found that for explicit attitudes for older people both the quantity and quality of contact were of influence. For implicit attitudes, only the quantity of contact was associated with more favorable implicit associations. Aberson and

Haag (2007) related their findings to the dual attitudes model (Wilson, Lindsey & Schooler, 2000) and specifically the MODE model (Fazio, 1990) mentioned above, which states that whether attitudes influence spontaneous or deliberative processes depends on the presence of the motivation and opportunity to affect explicit associations. Aberson and Haag (2007) noted that more or better contact might raise both implicit and explicit attitudes, which would reduce the role of the motivation and opportunity to change the explicit attitude as there will be less want or need to do so. This would make the explicit attitude (measured with the SRS) diverge less from the implicit attitude (measured by the IATs). This might explain why in the current study only the SRS was related with outgroup familiarity, but not the IATs. Higher outgroup familiarity would relatively alter explicit attitudes compared with implicit attitudes, which would explain the different correlations.

Despite this, it should be noted that the IATs still did not correlate with familiarity on their own. An explanation for this might be that familiarity was measured as a binary variable. If participants were familiar with either one or multiple people from the outgroup, they had to rate this familiarity. An issue with this method is that the amount of people participants know is not defined beyond either one or multiple people. This might be an important factor in defining exposure to the outgroup as both *more* and better contact might be of influence (Aberson & Haag, 2007), as a larger outgroup size can reduce anti-outgroup attitudes if this has led to more contact (Schlueter & Scheepers, 2010). A second issue with the design of the outgroup familiarity question method is that it cannot rule out the existence of secondary transfer effects of intergroup contact. In the current study participants were only asked about their experience with Moroccan people. Exposure to people from one outgroup can improve attitudes for other outgroups through a secondary transfer effect (Pettigrew, 2009; Tausch, Hewstone, Kenworthy & Psaltis, 2010; Harwood, Paolini, Joyce, Rubin and Arroyo, 2011). Thus, contact with other outgroups might have influenced the implicit and explicit attitude

results outside of contact with Moroccan people specifically. This transfer is more likely to occur if the two outgroups are more related or similar to one another (Harwood et al., 2011), so this would be mostly limited to outgroups that are to some extent similar to Moroccan people, such as Turkish people. In future studies on the effect of outgroup familiarity on P-IATs, these limitations should be taken into account.

Limitations

A selection of the above-mentioned influences on the W-IAT and P-IAT discrepancy double as limitations inherent in the study design for this study. Firstly, the representativeness of the P-IAT pictures has not been tested whilst the study used only two faces for the categories. Secondly, the affective stimuli were not homogenized and might have differed in salience. Thirdly, race and valence attributes were represented with face pictures and non-face pictures, which might have influenced the recoding process. Fourthly, the study did not account for the different levels of representation for the W-IAT (which used group labels) and P-IAT (which used faces of individuals). A running issue in these limitations is that the W-IAT and P-IAT were not fully equalized. In further studies, the crucial difference of using pictures only can be made clearer by taking these factors into account.

In addition to these factors, there is an important limitation that needs to be addressed as well: the study was conducted with an online questionnaire and thus lacked supervision of the input process. Furthermore, a potential issue lies in the automatic participant selection procedure. Participants had to indicate that they belong to the in-group for the study by virtue of the nationality and language of themselves and their parents after reading the introduction that already indicated the target group for the study. This selection was strict in order to minimize similarity with the outgroup. As it was impossible to know the extent of the influence of foreign cultures on each individual, participants could only participate if they and

both their parents were only of the Dutch nationality and spoke Dutch as their native language. This excluded all participants that belong to the target group of Dutch Caucasian despite minor influence from a foreign country. In addition, even western countries were excluded by this method due to cultural differences that might have influenced the results. Due to these limitations, the selection method should be kept in mind when generalizing the results.

Implications and further research

Although the results imply a discrepancy between the P-IAT and W-IAT, they nonetheless indicate that the P-IAT and W-IAT scores are (moderately) related. This implies that the P-IAT could be used as a possible alternative to word IATs. In its current form, this comes with the risk of yielding lower bias scores compared to W-IATs. When these scores are to be compared with previous word-based IATs, this is certainly an important point to consider for improvement. However, it should be noted that the P-IAT can tap into unique groups of participants by virtue of being a nonverbal task. The task can be useful for research with analphabetic participants, with younger children, and even with non-human primates. It is also useful for removing the effects of translation and language choice in cross-cultural research (although stimulus choice must still be considered carefully). Although the P-IAT can currently already be applied in these domains, more research is needed to standardize the P-IAT and thereby close the possible gap between the W-IAT and P-IAT scores.

In order to explore the potential of the P-IAT as an alternative test further, future studies could study the underlying mechanisms for the different scores between the P-IAT and W-IAT. As mentioned previously, unlike the W-IAT, the P-IAT has not been refined by rigorous testing and standardization. It would be interesting to use stimuli that are pre-tested for their representation of the whole category for the picture stimuli representing the categories. Furthermore, it would be interesting to test the use of positive and negative facial

expressions as a visual and socially relevant stimulus to represent the attribute of valence. Alternatively, further studies might study the benefit of raising the level of representation of the stimuli. They could do this by using pictures of groups of individuals clearly belonging to one social group, thereby building upon the study design used in an experiment by Foroni and Bel-Bahar (2010). As this equalizes the level of representation of the stimuli, this P-IAT design can be compared with W-IATs. Finally, future studies could explore the aforementioned target groups by using the P-IAT to circumvent the need for verbal processing. Although these studies would benefit from a more standardized P-IAT, their usage of an IAT would still provide useful results.

Acknowledgements

References

Aberson, C. L., & Haag, S. C. (2007). Contact, perspective taking, and anxiety as predictors of stereotype endorsement, explicit attitudes, and implicit attitudes. *Group Processes & Intergroup Relations, 10*(2), 179-201.

Aberson, C. L., Shoemaker, C., & Tomolillo, C. (2004). Implicit bias and contact: The role of interethnic friendships. *The Journal of Social Psychology, 144*(3), 335-347.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6, 10, and adulthood. *Psychological Science, 17*, 53-58.

Burns, J., Corno, L., & La Ferrara, E. (2015) Interaction, prejudice and performance. Evidence from South Africa. (*working paper*).

Chang, B. P. I., & Mitchell, C. J. (2011). Discriminating between the effects of valence and salience in the Implicit Association Test. *The Quarterly Journal of Experimental Psychology, 64*, 2251–2275. doi:10.1080/17470218.2011.586782

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development, 82*(3), 766-779.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4 year-olds: the preschool implicit association test. *Journal of Experimental Child Psychology, 109*(2), 187-200.

Danziger, S., & Ward, R. (2010). Language changes implicit associations between ethnic groups and evaluation in bilinguals. *Psychological Science, 21*(6), 799-800.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for white Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology, 36*, 316-328.

45

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50*(2), 77-85.

Fazio, R. H. (1990). Multiple Processes by which attitudes guide behaviour: The MODE model as an integrative framework. *Advances in Experimental Social Psychology, 23*, 75-109.

Foroni, F., & Bel-Bahar, T. (2010). Picture-IAT versus Word-IAT: Level of stimulus representation influences on the IAT. *European Journal of Social Psychology, 40*, 321-337.

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency. Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40*, 535-542.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216.

Greenwald, A. G., Poehlman, T. A., Uhlman, 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.

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.

Harwood, J., Paolini, S., Joyce, N., Rubin, M., & Arroyo, A. (2011). Secondary transfer effects from imagined contact: Group similarity affects the generalization gradient.

British Journal of Social Psychology, 50, 180-189.

Henry, P. J., & Sears, D. O. (2002). The Symbolic Racism 2000 Scale. *Political Psychology, 23*(2), 253-283.

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005a). What moderates implicit-explicit consistency? *European Review of Social Psychology, 16*(1), 335-390.

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

Jaeger, B. (2018). *Trait ratings for the Radboud Faces Database.* Retrieved from: https://psyarxiv.com/cf5ad/

Jones, R. K. & Hagen, M. A. (1980). Chapter 6 - A perspective on cross-cultural picture perception. In Jones, R. K. & Hagen, M. A. (eds), *Dürer's Devices: Beyond the Projective Model of Pictures* (193-226). Amsterdam: Elsevier.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America, 107*(25), 11163-11170.

Kubota, J. T., Peiso, J., Marcum, K., & Cloutier, J. (2017). Intergroup contact throughout the lifespan modulates implicit racial biases across perceivers' racial group. *PloS One.* DOI: https://doi.org/10.1371/journal.pone.0180440

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8.* University of Florida, Gainesville, FL.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24*(8), 1377-1388. DOI: 10.1080/02699930903485076

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49(*4), 764-766.

Meissner, F., Rothermund, K. (2013). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. *Journal of Personality and Social Psychology, 104*(1), 45-69. DOI: 10.1037/a0030734

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the Implicit Association Test. *Social Psychological and Personality Science, 6*(7), 740-748.

Möckel, T., Beste, C., & Wascher, E. (2015). The effects of time on task in response selection – An ERP study of mental fatigue. *Scientific Reports,* 5, 10113-10121. doi:10.1038/srep10113

Newheiser, A-K., Dunham, Y., Merril, A., Hoosain, L., & Olson, K. R. (2014). Preference for high status predicts implicit outgroup bias among children from low-status groups. *Developmental Psychology, 50*(4), 1081-1090.

Newheiser, A-K., & Olson, K. R. (2012). White and Black American children's implicit intergroup bias. *Journal of Experimental Social Psychology, 48*(1), 264-270.

Nisbeth, R., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134*(4), 565-584.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, 6, 101–115.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152-159.

Palfai, T. P., Kantner, C. K. & Tahaney, K. D. (2016). The image-based alcohol-action implicit association test. *Journal of Behavior Therapy and Experimental Psychiatry, 50*, 135-138. DOI: 10.1016/j.jbtep.2015.07.002

Paluck, E. L., Green, S. A., & Green, D. P. (2018). The contact hypothesis re-evaluated. *Behavioural Public Policy,* 1-30. DOI: 10.1017/bpp.2018.25

Park, B., Ryan, C. S., & Judd, C. M. (1992). Role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, 63, 553–567.

Pettigrew, T. F. (2009). Secondary transfer effect of contact: Do intergroup contact effects spread to noncontacted outgroups? *Social Psychology, 40*, 55–65.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Interpersonal Relations and Group Processes, 90*(5), 751-783.

Pieters, S. van der Vorst, H., Engels, R. C. & Wiers, R. W. (2010). Implicit and explicit cognitions related to alcohol use in children. *Addictive Behaviors, 35*(5), 471-478. doi: 10.1016/j.addbeh.2009.12.022

Qian, M. K., Heyman, G., Quinn, P. C., Messi, F. A., Fu, G., & Lee, K. (2016). Implicit racial biases in preschool children and adults from Asia and Africa. *Child Development, 89*(1), 1-12.

Qian, M. K., Heyman, G. D., Quinn, P. C., Fu, G., & Lee, K. (2017). When the majority becomes the minority: A longitudinal study of the effects of immersive experience

with racial out-group members on implicit and explicit racial biases. *Journal of Cross Cultural Psychology, 48*(6), 914-930.

Qian, M. K., Quinn, P. C., Heyman, G., Pascalis, O., Fu, G., & Lee, K. (2017a). A long-term effect of perceptual individuation training on reducing implicit racial bias in preschool children. *Child Development.* DOI: 10.1111/cdev.12971

Qian, M. K., Quinn, P. C., Heyman, G., Pascalis, O., Fu, G., & Lee, K. (2017b). Perceptual individuation training (but not mere exposure) reduces implicit racial bias in preschool children. *Developmental Psychology, 53*(5), 845-859.

Rutland, A., Cameron, L., Milne, A., & McGeorge, P. (2005). Social norms and self-presentation: Children's implicit and explicit intergroup attitudes. *Child Development, 76*, 451– 466.

Schlueter, E., & Scheepers, P. (2010). The relationship between outgroup size and anti-outgroup attitudes: A theoretical synthesis and empirical test of group threat- and intergroup contact theory. *Social Science Research, 39*, 285-295.

Setoh, P., Lee, K. J. J., Zhang, L., Qian, M. K., Quinn, P. C., Heyman, G. D., & Lee, K. (2019). Racial categorization predicts implicit racial bias in preschool children. *Child Development, 90*(1), 162-179.

Slabbinck, H., De Houwer, J., & Van Kenhove P. (2011). A pictorial attitude IAT as a measure of implicit motives. *European Journal of Personality, 25(*1), 76-86.

Smith, C. T., & Nosek, B. (2010). Implicit Association Test. In I. B. Weiner & W. E. Craighead (eds.), Corsini's Encyclopedia of Psychology, 4th edition. NY: Wiley & Sons Inc.

Steele, J. R., George, M., Williams, A., & Tay, E. (2018). A cross-cultural investigation of children's implicit attitudes toward White and Black racial outgroups. *Developmental Science, 21*(6). DOI: 10.1111/desc.12673

Tam, T., Hewstone, M., Harwood, J., & Voci, A. (2006). Intergroup contact and grandparent grandchild communication: The effect of self-disclosure on implicit and explicit biases against older people. *Group Processes & Intergroup Relations, 9*(3), 413-429.

Tausch, N., Hewstone, M., Kenworthy, J. B., & Psaltis, C. (2010). Secondary transfer effects of intergroup contact: Alternative accounts and underlying processes. *Journal of Personality and Social Psychology, 99*(2), 282-302.

Thomas, S. R., Smith, R. B., & Ball, P. J. (2007). Implicit attitudes in very young children: an adaptation of the IAT. *Current Research in Social Psychology, 13*(7), 75-85.

Turner, R., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology, 98*(3), 369-388.

Van Berlo, E., Otten, M., & Kret, M. (in prep). Measuring implicit associations with a pictorial IAT. (*unpublished*)

Williams, A., & Steele, J. R. (2016). The reliability of child-friendly race-attitude Implicit Association Tests. *Frontiers of Psychology.* DOI: https://doi.org/10.3389/fpsyg.2016.01576

Wittenbrink, B., Judd, C. M, & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827.

Appendix

Appendix A - Methods Supplementation

Table 1:

The randomization of starting positions for each block of an IAT. Each participant will only complete one randomly assigned version for each IAT. M = Moroccan race concept, C = Caucasian race concept, P = positive attribute and N = negative attribute.

Block	Type	Trials	Tested	Version 1	Version 2	Version 3	Version 4
1	Training	20	Concepts only	M - C	M - C	C - M	C - M
2	Training	20	Attributes only	N - P	P - N	P - N	N – P
3	Experimental	20	Combined	MN - PC	MP - NC	CP - NM	CN - PM
4	Experimental	20	Combined	NM - CP	PM - CN	PC - MN	NC - MP
5	Training	20	Attributes only	P - N	N - P	N - P	P - N
6	Experimental	20	Combined	MP - NC	MN - PC	CN - PM	CP - NM
7	Experimental	20	Combined	PM - CN	NM - CP	NC - MP	PC - MN

Appendix B – Stimuli in P-IAT

The P-IAT made use of the RaFD pictures shown in figure 14 and IAPS pictures shown in figure 15.
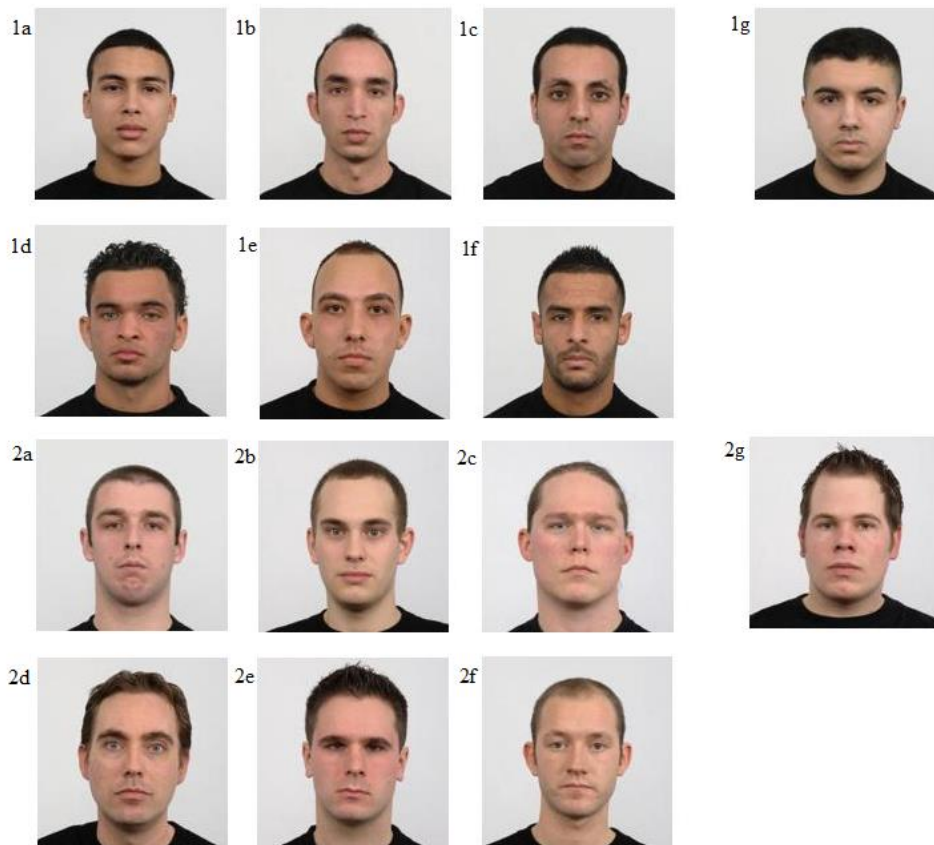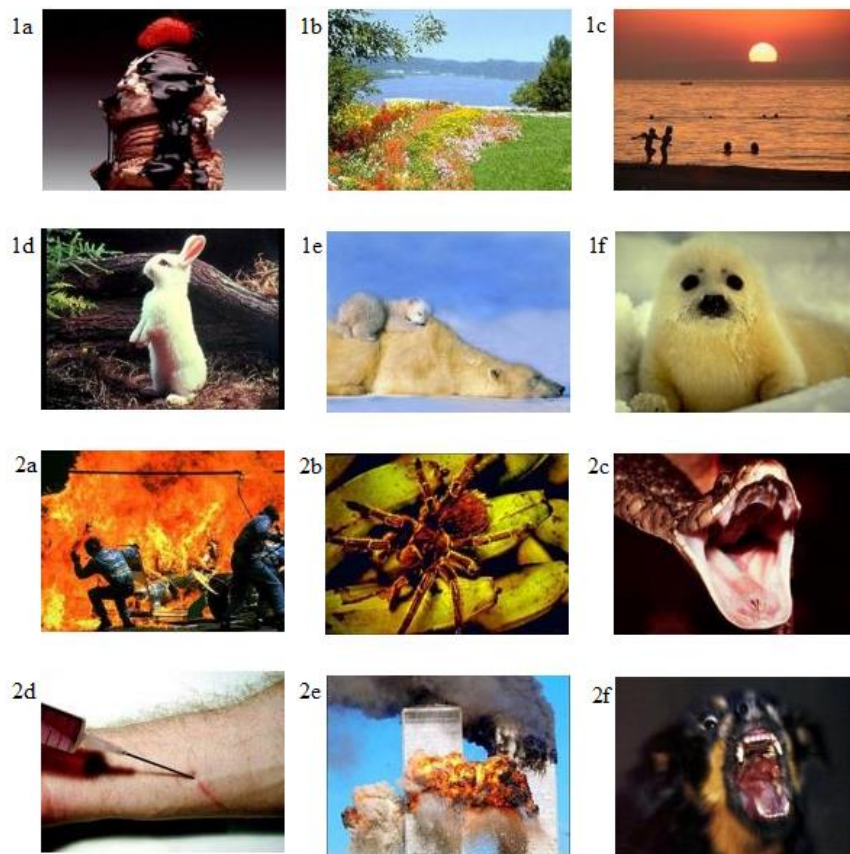
Figure 14. The RaFD pictures used in the P-IAT task (category and target) and W-IAT task (targets only). Picture 1a-1f were exclusively used as stimuli to represent Moroccans. Picture 1g was exclusively used to represent the category 'Moroccan' in the P-IAT task. Picture 2a-2f were exclusively used as stimuli to represent Caucasians. Picture 2g was exclusively used to represent the category 'Caucasian' in the P-IAT task.

Figure 15. The IAPS pictures used in the P-IAT task (category and target) and W-IAT task (targets only). Picture 1a-1e were exclusively used as stimuli to represent the attribute 'positive'. Picture 1f was exclusively used to represent the category 'positive' in the P-IAT task. Picture 2a-2e were exclusively used as stimuli to represent the attribute 'negative'. Picture 2f was exclusively used to represent the category 'negative' in the P-IAT task.

Appendix C – Symbolic Racism Scale (SRS), Dutch translation.

1. Het is echt een kwestie van niet hard genoeg werken; als Marokkanen harder zouden werken dan zouden ze het even goed hebben als Nederlanders.
- 1. Volledig mee eens
- 2. Gedeeltelijk mee eens
- 3. Gedeeltelijk mee oneens
- 4. Volledig mee oneens

2. Surinamers, Polen en andere minderheidsgroepen zijn over de vooroordelen heen gekomen en hebben zichzelf omhoog gewerkt. Marokkanen zouden hetzelfde moeten doen.

- 1. Volledig mee eens
- 2. Gedeeltelijk mee eens
- 3. Gedeeltelijk mee oneens
- 4. Volledig mee oneens

3. Sommige mensen vinden dat Marokkanen te hard hun eigen normen en waarden door proberen te drukken. Anderen vinden dat Marokkanen niet goed genoeg voor hun eigen normen en waarden opkomen. Wat vind jij?

- 1. Marokkanen proberen te hard hun eigen normen en waarden door te drukken
- 2. Marokkanen moeten er meer voor opkomen
- 3. Marokkanen komen niet teveel en niet te weinig op voor hun eigen normen en waarden

4. In Nederland heersen er veel spanningen m.b.t. minderheden en vreemdelingen. Hoe erg zijn Marokkanen volgen jou verantwoordelijk voor het ontstaan van spanningen?

- 1. Volledig verantwoordelijk
- 2. Grotendeels verantwoordelijk
- 3. Een beetje verantwoordelijk
- 4. Niet erg verantwoordelijk

5. Hoeveel wordt er tegen Marokkanen gediscrimineerd, vind jij?

- 1. Erg veel
- 2. Redelijk veel
- 3. Een beetje
- 4. Helemaal niet

6. Het discrimineren van minderheden zoals Marokkanen zorgt ervoor dat deze groepen moeilijk uit lagere sociale milieus kunnen komen.

- 1. Volledig mee eens
- 2. Gedeeltelijk mee eens
- 3. Gedeeltelijk mee oneens
- 4. Volledig mee oneens

7. Marokkanen hebben de laatste jaren veel minder gekregen dan dat ze eigenlijk verdienen.

- 1. Volledig mee eens
- 2. Gedeeltelijk mee eens
- 3. Gedeeltelijk mee oneens
- 4. Volledig mee oneens

8. Marokkanen hebben het de laatste jaren financieel veel beter gehad dan dat ze eigenlijk verdienen.

- 1. Volledig mee eens
- 2. Gedeeltelijk mee eens
- 3. Gedeeltelijk mee oneens
- 4. Volledig mee oneens

Appendix D – Outlier removal graphs

When the median absolute deviation (MAD) method was used, there were three possible levels of level of conservatism. We chose the moderately conservative level based on the frequency distributions (figure 16) and Q-Q plots (figure 17) of the W-IAT and P-IAT D-Scores after the outliers for these criteria were removed.
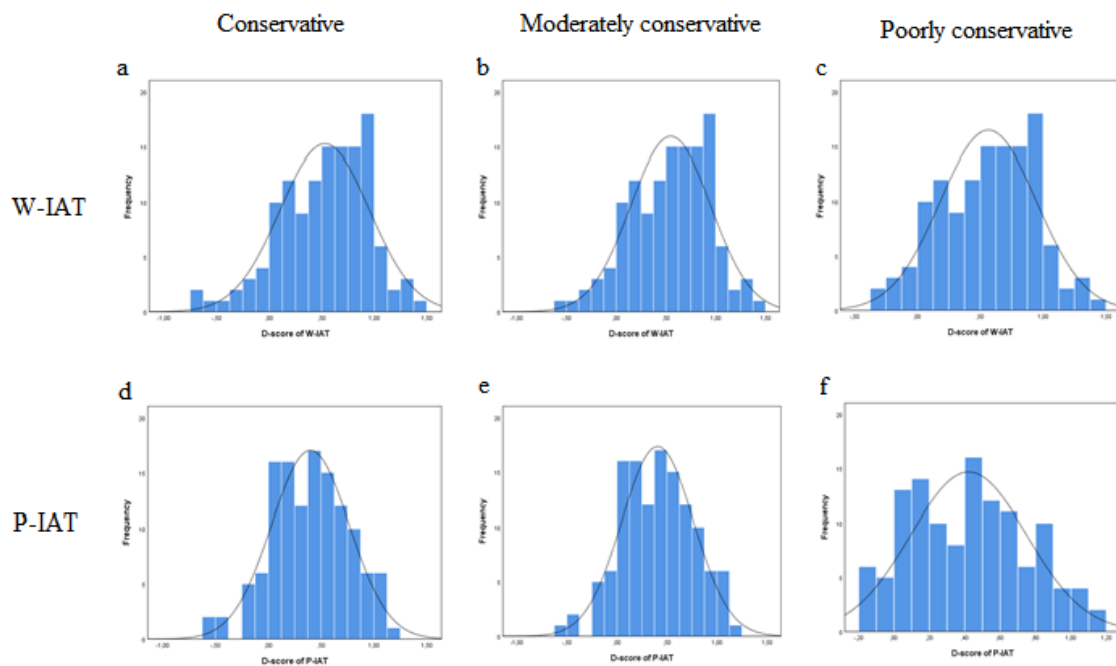

Figure 16. Frequency distributions of W-IAT (a-c) and P-IAT (d-f) scores after outlier removal for each level of conservatism of the MAD method.
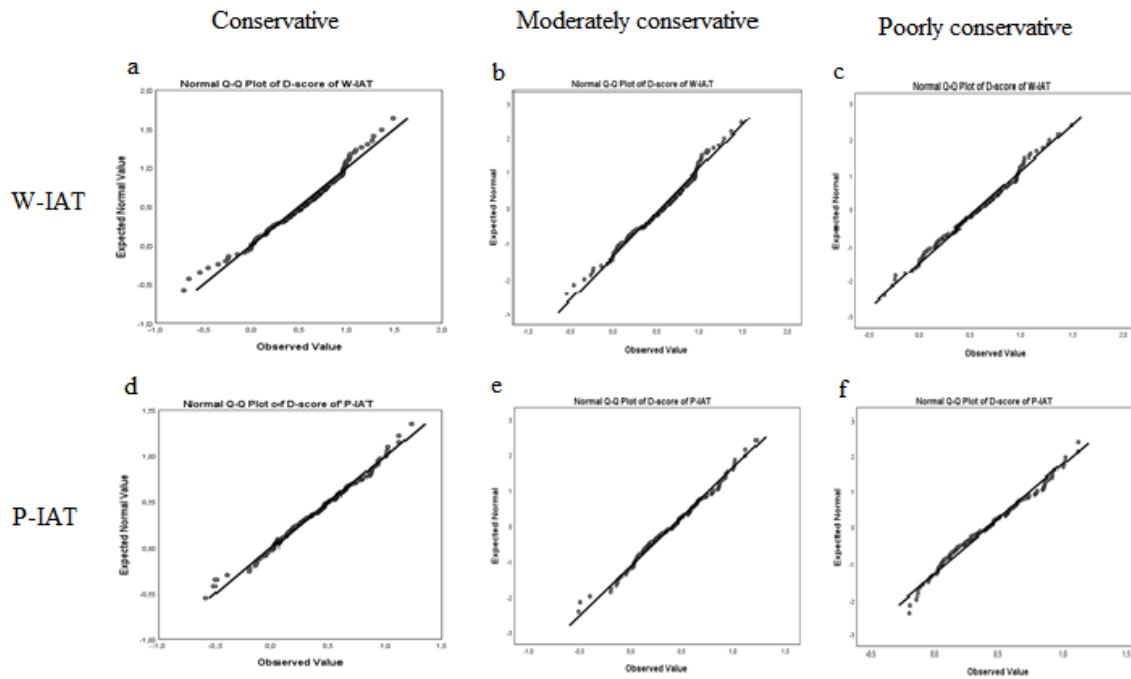
Figure 17. Q-Q plots of W-IAT (a-c) and P-IAT (d-f) scores after outlier

removal for each level of conservatism of the MAD method.