Emotions through the eyes of our closest living relatives: exploring attentional and behavioral mechanisms

Berlo, E. van

Part III:

Implicit Associations

Chapter 7

Validation of a pictorial version of the Implicit Association Test (IAT) for comparative research

Abstract

The Implicit Association Test (IAT) is frequently used to measure implicit associations, but one issue with current forms of the IAT is that they require at least some comprehension of written and/or spoken language, making them difficult to use in illiterate or pre-verbal populations. We therefore designed a self-explanatory, touchscreen-based, fully pictorial IAT that we validated across two experiments. In Experiment 1 we tested the PIAT's ability to measure implicit inter-ethnic attitudes in 129 Dutch adults and in 143 Dutch children visiting a zoo. In Experiment 2, we validated the PIAT by directly comparing its results to a word IAT in an online, within-subjects experiment involving 141 adults. D-score analyses showed that the PIAT can reliably tap into the same implicit biases as its verbal counterpart. We believe that the PIAT provides a good adaptation to the original IAT, offering a standardized test that could potentially be suitable for quantitative, cross-cultural, and cross-species comparisons.

Based on:
Van Berlo, E., Otten, M., Roth, T. S., Binnekamp, J., Van der Ven, E. J., & Kret, M. E. (2021). Validation of a pictorial version of the IAT. *Manuscript submitted for publication*.

Introduction

Since the publication of the seminal paper by Greenwald, McGhee, and Schwartz (1998), the Implicit Association Test (IAT) has been one of the most well-established tasks to measure implicit attitudes (Cunningham et al., 2001; Greenwald et al., 2003, 2009; Lane et al., 2007). Implicit tasks such as the IAT are crucial for providing a window into unconscious processes that drive behavior. Unfortunately, most versions of the IAT require comprehension of written or spoken language, thereby limiting its usability in non-verbal or illiterate populations such as young children, clinical populations such as individuals on the autism spectrum or cognitively impaired individuals, and possibly in comparative research with non-human animals as well. To offer researchers who work with these populations an adaptation to verbal IATs, we developed a fully pictorial, intuitive implicit association test, and validated it against its classical counterpart in a population of children and adults.

The original IAT has received extensive psychometric evaluation and is a widely used tool to assess implicit attitudes and stereotypes (Cunningham et al., 2001; Greenwald et al., 2003, 2009; Lane et al., 2007). It measures the strength of implicit associations between two concepts (e.g., names of African-Americans vs. White Americans) and two attribute dimensions (e.g., pleasant vs. unpleasant words) by comparing reaction times in a categorization task consisting of a series of testing blocks. In the practice blocks, participants learn to categorize exemplars of the concepts and attributes into their superordinate categories. For example, they categorize a name such as "Tyrone" as African-American and "Hannah" as White-American, and a word such as "happiness" as pleasant and "suffering" as unpleasant. In the critical blocks, these superordinate categories are combined. For instance, in the first critical block participants categorize names and faces into the combined superordinate categories "White" + "unpleasant" and "Black" + "pleasant". In the subsequent critical block(s), this combination is reversed (e.g., "White" + "unpleasant"). In general, participants respond faster in critical blocks congruent with their implicit associations, and slower in incongruent critical blocks (Greenwald et al., 1998). Importantly, in order to complete a typical IAT, a good understanding of written and/or spoken language is necessary, as the task activates implicit attitudes through the use of words and names representing (a subset of) the superordinate categories.

Several studies have made elegant adaptations to the IAT to partly overcome the necessity for understanding written or spoken language, especially within the developmental sciences. For instance, one IAT used pictures of flowers and insects

to represent the concept categories, but retained words for describing the attribute dimensions (e.g., "good" and "bad") in order to test implicit attitudes in 6- and 10-year-olds (Baron & Banaji, 2006, but also see a pictorial and touchscreen adaptation by Thomas et al., 2007). In the pre-school IAT, Cvencek and colleagues (2011) used schematic representations of smiling and sad faces to indicate the attribute dimensions "good" and "bad", and pictures of flowers and insects for the concepts. Furthermore, during the task pictures and (spoken) words were alternated (Cvencek et al., 2011). In another child-friendly IAT, only pictures of White and Black faces were used, together with line drawings of happy and sad faces (Rutland et al., 2005). Instead of pushing buttons that were mapped with the left and right superordinate categories, children had to move their mouse towards the target locations. Using a touchscreen, another adaptation of the IAT used pictures of faces that were shown in the middle of the screen with a happy and sad emoticon on the left and right side on the bottom of the screen. In critical blocks, children were told whether to press the "happy" or "sad" emoticon (Qian et al., 2016; Setoh et al., 2019). Notwithstanding the high usefulness of these tasks, a commonality in the adaptations is that child participants still received (extensive) instructions (for instance, extra verbal instruction during critical blocks (Qian et al., 2016; Setoh et al., 2019)), or the tasks combined words with pictures (Baron & Banaji, 2006; Cvencek et al., 2011). To this end, we focused on creating an entirely non-verbal pictorial version of the IAT (from hereon: PIAT) with stimuli that are proven to be interpreted similarly across cultures, and that requires a minimal amount of instructions to complete.

The aim of our study was to assess our PIAT's validity in comparison to its classic counterpart. In Experiment 1, we test the PIAT in a large, diverse population consisting of Dutch adults and children that are visiting a zoo with two aims in mind, i) to assess whether our version of the PIAT can indeed measure implicit attitudes, and ii) to assess whether it is suitable for use in a heterogeneous sample in a more naturalistic environment (i.e., not in the lab). In Experiment 2, we directly compare the performance of an online version of our PIAT to an online word IAT (WIAT) with the aim to further validate it as a tool for measuring implicit attitudes. We have chosen to measure implicit attitudes towards different ethnicities, as these have also been extensively studied in traditional IATs (Baron & Banaji, 2006; Greenwald et al., 1998). In both experiments, we assess participants' implicit attitudes towards individuals of Moroccan and Dutch descent using images from validated face and emotional images databases (Lang et al., 2007; Langner et al., 2010). We chose for these ethnicities because negative opinions about individuals of Moroccan descent

are pervasive in Dutch society (Andriessen et al., 2020). For instance, compared to other minorities, individuals of Moroccan descent were rated as most negative in a Dutch national survey (Vrooman et al., 2014). As such, we expect to find that Dutch participants (i.e., participants with 2 parents born in the Netherlands) have negative implicit associations with individuals of Moroccan descent, compared with individuals of Dutch descent. Furthermore, we expect to find this bias in both adults and children. Finally, we expect that the results of the PIAT are comparable to those found in word-based IATs.

Experiment 1: PIAT in the zoo

Method

Participants

129 Adults (73 females) and 143 children (72 females) took part in this study. Participants were visitors of a zoo in the Netherlands, and were all Dutch-speaking and of Dutch descent. The majority of individuals were right-handed (adults: 113 participants (87.6%); children: 121 (84.6%)). Children were between ages 5 to 17 (M = 10, SD = 2.29) and adults between ages 18-76 (M = 33.83, SD = 13.95). Consent for participating in the study was received from all adult participants and parents of participating children. Participants took part in the study on a voluntary basis and thus were not compensated for their participation. Data were collected between April and May 2017 (see Figure S1 in supplements for photos of the setup).

Task

The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) measures the strength of an association between a pair of concepts based on the reaction times of the participant. In a series of blocks, participants categorize concepts into two opposing categories. In the original version, seven blocks of trials are used, but versions with five blocks are common as well, especially when the target group consists mainly of children (Nosek et al., 2005). We thus opted for the five-block version of the task. Block one, two, and four were practice blocks in which participants learned to categorize images; block three and five were critical blocks in which the speed of categorization was measured (Figure 1).

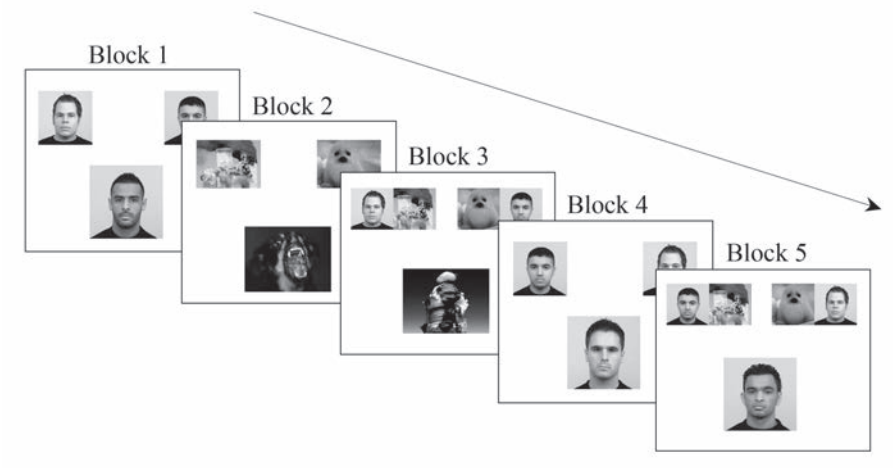

Figure 1. Block design of the PIAT. Here, the adult version of the PIAT is shown. Block 1 and 2 always consisted of trials in which participants had to categorize faces and positive and negative images (in random order). Block 3 is the first critical block in which concepts and attributes are combined. In Block 4, the position of the faces is reversed. Block 5 is the last critical block.

Practice blocks consisted of 20 trials, and critical blocks included 40 trials. Before the start of each trial, a "fixation" dot appeared in the lower-middle part of the screen, which had to be touched in order to start or continue the task (Figure 2A, I). The dot functioned as a fixation cross that directed the participants' gaze and hand towards it, thereby preventing an attentional bias to the left or right side of the screen.

In the first practice block, participants learned to categorize images into superordinate categories (i.e., faces were categorized into the superordinate "Moroccan" and "Dutch" categories, each represented by one exemplar image of a man of Moroccan or Dutch descent). These concepts were presented on the top left and right side of the screen, while the images that had to be categorized were presented in the lower-middle part of the screen. Correct categories could be indicated by pressing on the exemplar image on the screen (Figure 2A, II), and feedback was given in the form of a thumbs-up or -down image[2] (2s), indicating a correct or incorrect answer (Figure 2A, III). Next, a dot appeared that had to be touched to start the next trial (Figure 2A, IV-V). In the second practice block of the PIAT, participants again categorized images into two categories, but this time the images represented a positive or negative attribute dimension. In block three (the first critical block), the

2 Populations unfamiliar to this sign will in a separate learning phase need to learn that thumbs up means 'good' and down, 'bad'. Alternatively, different culture-specific signs can be implemented instead.

concepts from block one and attribute dimensions of block two were combined (Figure 2B, and see also Figure S2A in the supplements). In practice block four, the spatial location of the concepts (Moroccan or Dutch descent) was switched, so that the concept that was on the right side in block 1 was now on the left side, and vice versa. Block five was similar to block three and formed the second critical block, but

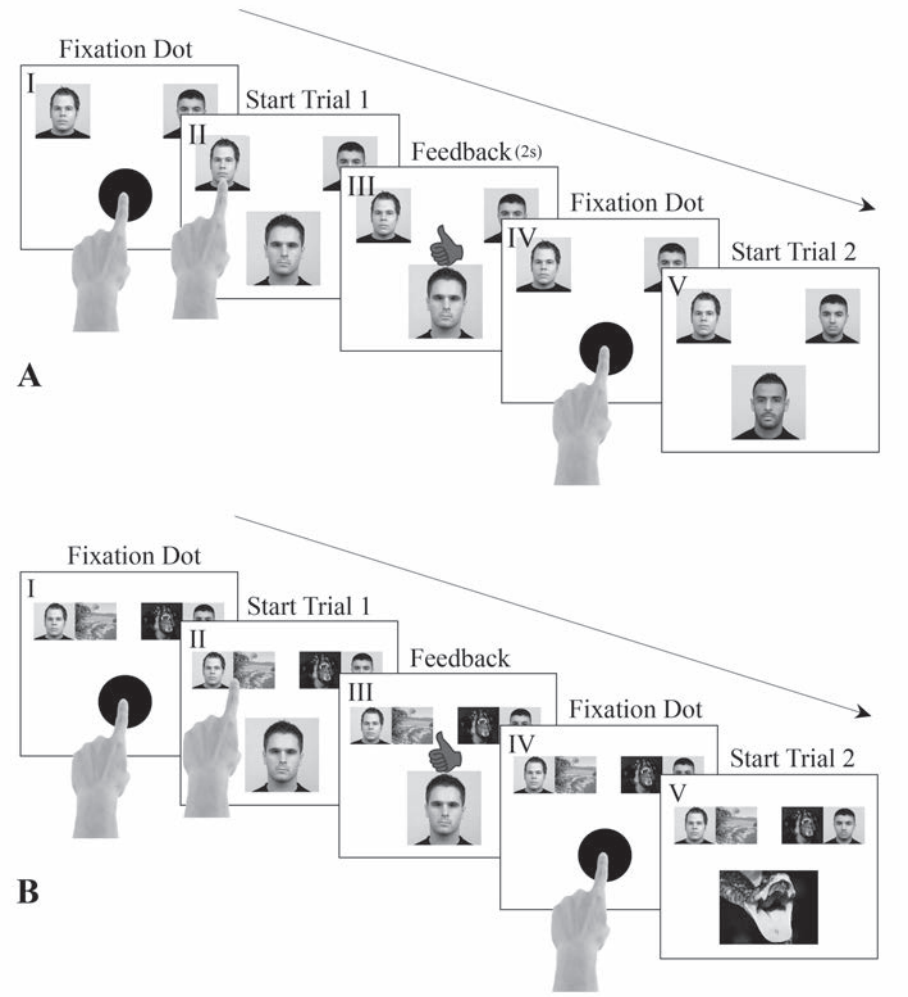

Figure 2. (A) Trial outline for a practice block in the PIAT. After pressing the dot (I), participants categorize the image they are presented with by tapping on the correct concept (faces), attribute dimension (positive/negative scenes), or a combination of the two appearing in the top left and right corner of the screen (II). Participants then receive feedback in the form of a thumbs-up or -down (III, presented for 2 s). Next, a new dot appears in order to start the next trial (IV, V). (B) Trial outline for the critical blocks of the adult PIAT. The trials in the critical blocks follow the same procedure as in the other blocks. In the child PIAT, attribute images are replaced by cartoon figures.

this time the concept-attribute combination was switched (i.e., if participants saw the face of a man of Moroccan descent combined with a negative image in block three, they now saw the face of a man of Moroccan descent combined with a positive image, Figure 2A and 2B).

The IAT is known for its order effects such that when congruent trials are presented first in the critical block, larger IAT effects are found than when incongruent trials are presented first (Nosek et al., 2005). Order effects can be counteracted by counterbalancing the presentation order of the critical blocks, and therefore half of the participants started out with congruent trials (i.e., "Dutch" + positive, "Moroccan" + negative), while the other half started out with incongruent trials (i.e., "Dutch" + negative, "Moroccan" + positive). Furthermore, in the practice blocks, participants either started out with categorizing faces (concepts) or with categorizing positive and negative scenes (attributes). Reaction times on trials in the critical blocks were expected to depend on the congruency of the trials, i.e., following our hypothesis, participants were expected to respond slower to incongruent trials in which faces of men of Dutch descent were linked to negative attributes, and the faces of men of Moroccan descent to positive attributes.

The PIAT was performed on a touchscreen using only one hand. This is different from the typical procedure where participants use their left and right hand to press a left and right key on a keyboard. Previous studies have shown that handedness and the assignments of the left or right response key to a particular IAT category have little to no influence on the IAT effect (Greenwald, 2001; Greenwald et al., 1998), thus we expect no difficulties with using one hand in the PIAT.

Stimuli[3]

Adult

In non-verbal versions of the IAT, stimuli usually consist of pleasant and unpleasant words (attributes) and words referring to the two categories that are being investigated (concepts). In the PIAT, concepts and attributes were replaced with images. Concepts

3 Given the sensitivity of topics such as ethnic prejudice and discrimination, I wish to provide some more context to our choice for measuring implicit racial attitudes (and not, for instance, implicit attitudes towards insects, established in prior studies (Greenwald et al., 1998)). Originally, when designing the study, we aimed to validate it in humans and subsequently test bonobos and chimpanzees on their implicit associations with familiar group mates (the so called "ingroup") and unfamiliar others ("outgroups"). By doing so, we wanted to gain more insights into the evolutionary roots of discrimination. We therefore chose to test the implicit association that humans may have with their ingroup (e.g., individuals of the same descent) versus one potential outgroup (e.g., individuals of Moroccan descent).

consisted of six images of faces of men of Moroccan or Dutch descent with a neutral expression and formed a subset of the Radboud Faces Database (Langner et al., 2010) (see Table S1.1 and Figure S2A in the supplements). For the attributes, we selected the six most negatively and most positively rated images from the International Affective Picture System (Lang et al., 2007), excluding images showing humans as those may interfere with measuring inter-ethnic attitudes (Table S1.2 and Figure S2B in supplements). All images were presented in color, and all images were presented twice during the two critical blocks.

To control for an effect of attribute image type on the results, three different stimulus sets were created (Figure S3A in supplements). In all sets, we used the same faces to depict the "Moroccan" and "Dutch" superordinate category, but varied the images depicting the positive and negative category. For example, in the first version of the stimulus set, the positive attribute was represented by a seal pup and the negative attribute by a building on fire. For each stimulus set, we created four different versions in order to control for order effects and stimulus location effects within the task.

Images that served as category indicators were 300x300 pixels and were presented in the top-left and top-right corner of the screen. In the experimental trials (block 3 and 5, but see *Task*) a combined image of a face and a positive or negative attribute was shown on the top-left and top-right corner of the screen with a dimension of 500x225 pixels (see Figure 1B). Finally, images that needed to be categorized appeared in the lower-center part of the screen and were 400x300 pixels (attributes, i.e., positive/negative scenes) or 450x450 pixels (concepts, i.e., faces).

Children

The faces used as stimuli in the adult PIAT were also used in the child PIAT. Some IAPS images can be upsetting or frightening for young individuals, thus we opted to use cartoons instead. Positive and negative attribute images were changed to cartoon heroes and villains from animated tv-shows (see Table S1.3 and Figure S2C in the supplements for an overview of the selected images). All these attribute images were rated based on valence and arousal by children, and thus validated before we commenced the study (see Table S1.3 in supplements for an overview of the results). Like in the adult PIAT, we used six different positive and negative attributes, each consisting of a hero or villain from the same cartoon (Figure S2C in supplements). The child PIAT followed the same procedure and task presentation as the adult PIAT (i.e., the presentation order between practice and critical blocks were counterbalanced).

Furthermore, like in the adult PIAT, three different stimulus sets were created in which the positive and negative attribute images were varied. Again, for each stimulus set, we created four versions to control for order effects (see Table S2 in supplements and Figure S3B).

Equipment

The adult PIAT was performed on a Dell S2240Tb touchscreen (21.5 inch, 1920x1080 pixels, 12 ms response time). Children performed the PIAT on an Iiyama T1931SR-B1 touchscreen (19 inch, 1280x1024 pixels, 5 ms response time). Validation of the child PIAT attribute images was conducted on a Panasonic FZ-G1 ToughPad tablet (10.1 inch, 1920x1200 pixels).

Procedure

Participants were recruited by student assistants who approached zoo visitors that passed the test location during their visit. The assistants approached the visitors with information about the studies taking place in the zoo, and visitors were then asked if they were willing to participate in the current study. The goals of the study were deliberately kept vague and only minimal instructions were provided. Participants were told that the current test required them to categorize the big picture (of either a face or scene) into one of two categories on the upper left or right side of the screen, and that the test itself would provide them with feedback on their performance. They were also instructed to only use only one hand. After receiving consent from adult participants and the caregiver of child participants, individuals were seated behind the touchscreen. Participants started out with five practice trials in which they sorted images of flowers and bunnies to get a better idea of how the task looks and works. Next, they completed the five blocks of the PIAT. The task took about 10 minutes to complete. After completion, participants were thanked for their participation and fully debriefed.

Analyses

Analyses were performed in R, using the IATscores package (Richetin et al., 2015). We calculated a D-score using RobustScores (a function within IATscores) based on the following minimum performance criteria: reaction lower than 10,000 ms, and an error rate below 40% for the critical blocks (Greenwald et al., 2003; Nosek et al., 2014). Furthermore, in one of the stimulus sets used in 15 children (Stimulus Set 2, Version 2), one of the trials in critical block 3 wrongly presented participants with

superordinate category images that belonged to the last critical block (block 5). For each of the affected participants, the reaction time of the trial was manually set to 10,001 ms to ensure that it would be discarded in subsequent analyses.

For adults, two trials were removed due to the 10,000 ms cutoff, and one participant was excluded due to a high error rate (Mean error in our sample M_{error} = 4.91, SD = 6.31). For children, 44 Trials were excluded based on the 10,000 ms criterion (15 trials within one child; the rest divided over 29 children), and one child was excluded based on the 40% error cutoff (M_{error} = 6.92, SD = 7.23). Remaining reaction time data were then 10% winsorized (Richetin et al., 2015). Note that we did not exclude erroneous trials; inclusion of erroneous trials increases validity and reliability of the scoring method (Richetin et al., 2015). The D-score represents the difference in reaction times (after processing) between critical blocks divided by the standard deviation of the datapoints in both critical blocks. Positive D-scores indicated an association between faces of men of Moroccan descent and negative images, and faces of men of Dutch descent and positive images.

For adults and children we performed two separate one sample t-tests to establish whether D-scores significantly differed from 0. Furthermore, as presentation order of the critical blocks and task version may affect D-scores, we also fit two separate linear models using sum-to-zero coding. We used *Congruency* (i.e., congruent or incongruent block first) and *Task Version* (Version 1, 2 or 3, reflecting the three different stimulus sets) as fixed effects, with the intercept reflecting the average D-score in our sample. Next, to assess the consistency of results across all items, we correlated results of the first half of the trials within the critical blocks with the second half of the trials using the SplitHalf function in IATScores.

Furthermore, the results of sensitivity power analyses for our main hypotheses can be found in the supplements (supplemental Figure S5).

Results

Adult PIAT

For adults we found a significantly positive D-score average of .24 (*95% CI* [.16, .32], $t(127) = 5.94$, $p < .001$, *Cohen's d* = .52), indicating that adults respond significantly faster on congruent trials ("Moroccan" + negative and "Dutch" + positive) versus incongruent trials ("Moroccan" + positive and "Dutch" + negative; Figure 3 and Table 1). After controlling for *Congruency* and *Task Version*, this effect remained present

(D-score: .24, *95% CI* [.16, .32], $t(124) = 5.97$, $p < .001$). We also found a main effect of order ($F(1, 124) = 14.22$, $p < .001$), meaning that adults who receive incongruent trials first show a higher D-score on average ($M = .39$, $SD = .41$) than adults who receive congruent trials first ($M = .09$, $SD = .46$). We found no main effect of task version on D-score ($F(2, 124) = .07$, $p = .936$, Table 1 as well as supplemental Figure S4). Lastly, when assessing the internal consistency of the adult PIAT, we found split-half reliability of $r = .84$.

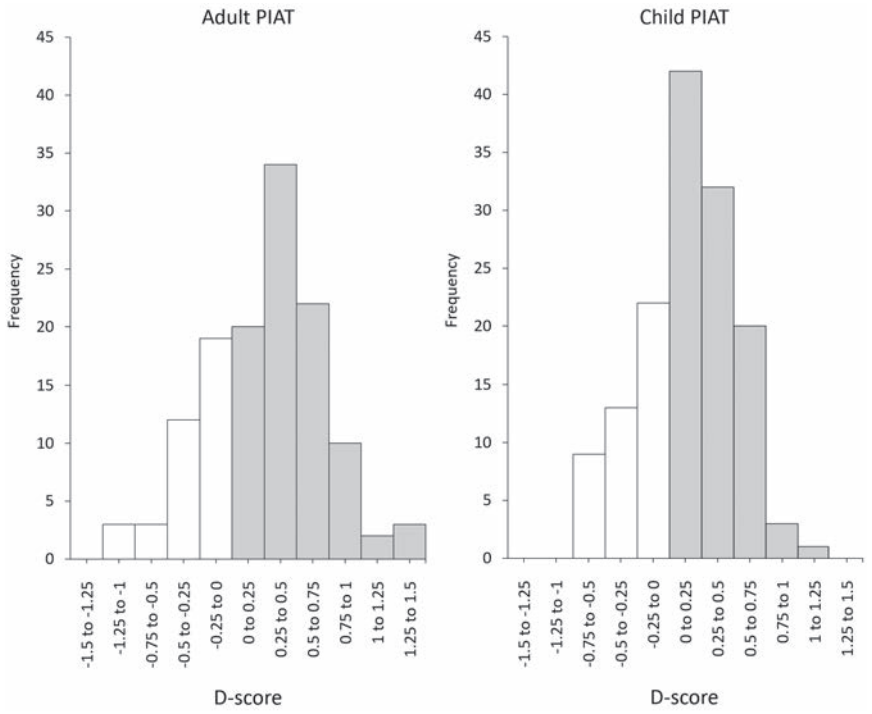

Figure 3. D-score distribution for the adult (left) and child PIAT (right). Positive values represent stronger associations between pictures of men of Dutch descent and positive scenes, and pictures of men of Moroccan descent and negative scenes.

Child PIAT

Children showed statistically significant positive D-score average of .14 (*95% CI* [.08, .21], $t(141) = 4.56$, $p < .001$, *Cohen's d* = .38), meaning that children associate faces of men of Dutch descent with positivity, and faces of men of Moroccan descent with negativity (Figure 3, Table 1). Using a linear model with *Congruency* and *Task Version*

added, this effect remained present (D-score: .15, *95% CI* [.09, .21], $t(138) = 4.83$, $p < .001$). We also found a significant effect of *Congruency* on D-score ($F(1, 138) = 9.18$, $p = .003$); children who received incongruent trials first responded significantly faster than children who received congruent trials first (incongruent: $M = .25$, $SD = .34$; congruent: $M = .06$, $SD = .39$, supplemental Figure S4). Additionally, there was a significant effect of *Task Version* on D-scores ($F(2, 138) = 3.08$, $p = .05$, Figure S4). When plotting the data, all task versions cause the IAT effect in the expected direction (i.e., an average D-score above 0), but task version 1 causes the lowest ($M = .05$, $SD = .38$) and task version 3 the highest effect ($M = .18$, SD = .34, task version 2: $M = .17$, $SD = .40$, S3). Indeed, post-hoc comparisons using the Tukey procedure revealed a significant difference between task version 1 and 3 (V1-3: $p = .040$, V1-2: $p = .233$, V2-3: $p = .627$).

Finally, to assess the consistency of results across all items, we correlated D-scores based on only the first half of the trials within the critical blocks with D-scores based on only the second half and found a split-half reliability of $r = .69$, indicating acceptable internal consistency in the child PIAT.

Table 1. Model results for the adult and child PIAT

				D-score		
Adult PIAT	*Predictors*	*Estimates*	*SE*	*95% CI*	*t*	*p*
	(Intercept)	.24	.04	.16 – .32	5.97	**<.001**
	Congruency (Congruent first)	.15	.04	.07 – .23	3.78	**<.001**
	Task Version (V2)	-.01	.06	-.12 – .11	-.13	.900
	Task Version (V3)	-.01	.06	-.12 – .10	-.21	.832
Child PIAT	(Intercept)	.15	.06	.09 – 0.21	4.83	**< .001**
	Congruency (Congruent first)	-.21	.06	.04 – 0.17	3.37	**.001**
	Task Version (V2)	.12	.08	-.19 – .02	-2.33	**.02**
	Task Version (V3)	.19	.08	-.06 – .10	.45	.65

Conclusion

The results show that both children and adults appeared to have a significantly negative implicit bias towards faces of men of Moroccan descent and a positive implicit bias towards faces of men of Dutch descent, but children appear to have a weaker implicit bias than do adults. The PIAT thus appears suitable for testing both

adults and children in an environment with a lot of potential distractors (i.e., the zoo). Furthermore, in adults the different versions of the tasks did not significantly impact D-scores, showing that the specific stimuli used in the PIAT do not significantly impact the PIAT's ability to measure implicit attitudes, at least when using the IAPS and Radboud Faces database images. In the child PIAT we do find a significant effect of task version 1 on D-scores. Finally, the PIAT shows decent internal consistency with results that are in line with previous findings in image-based IATs (e.g., split-half reliability $r = .69$ in Palfai et al., 2016).

The IAT can suffer from order effects (Nosek et al., 2005) and despite counterbalancing critical block order, this is also apparent in our PIAT; Individuals who receive incongruent trials first show a much higher D-score average than individuals who receive congruent trials first. This contrasts with common IAT findings that show the reverse effect, i.e., higher IAT effects are found for tasks that present the congruent trials first (Nosek et al., 2005).

To further validate the PIAT, in the next experiment, we investigate whether the effects we find in the PIAT correlate with results on an IAT combining words and pictures (word-IAT or WIAT) within the same subjects, and also study whether both IATs are correlated with more explicit measures of inter-ethnic biases.

Experiment 2: Online PIAT and WIAT

Method

Participants

Initially, 158 adult participants took part in the online PIAT/WIAT study, but 17 did not complete the study, thus resulting in a final N of 141 (Age range: 19-68, $M = 23.72$, $SD = 10.16$, 114 females). All participants were native Dutch speakers with a Dutch nationality, and had parents with the Dutch nationality as well. Most of the participants were right-handed (i.e., 128 (81%)). All participants were recruited via an online recruitment system of Leiden University (SONA), through flyers and posters, and through social media. As part of the Psychology curriculum of Leiden University, participants received 1 course credit after completing the experiment. Data collection took place between June 2018 and March 2019.

Tasks

Participants took part in two 7-block IATs; a word-IAT in which individuals categorized pictures into categories represented with words (WIAT) and our picture-only IAT (PIAT). In both IATs, there were initially two training blocks (block 1 and 2), each consisting of 20 trials. Next, participants continued through two critical blocks (block 3: 20 trials, block 4: 40 trials), followed by another training block (block 5: 40 trials). Finally, participants completed two critical blocks again (block 6: 20 trials, block 7: 40 trials). In contrast with Experiment 1 that was conducted on touchscreens, participants now performed the tasks online while using a keyboard. Participants used the "E" and "I" keys to indicate the left and right superordinate categories of the concepts (faces) and attributes (positive and negative images), respectively, and we used their reaction time on the key presses as a measure of bias. Furthermore, whereas on the touchscreen-based task participants had to press a dot to continue, participants were now shown a fixation cross in the middle of the screen for 300 ms before the next trial started.

The task design for the WIAT and PIAT was similar to Experiment 1, but the two critical blocks each contained an extra 20 trials. In the first training block, participants categorized perceived ethnic concepts ("Moroccan" vs. "Dutch"), and in the second training block attributes (positive vs. negative). The third and fourth blocks, i.e., critical blocks, presented participants with the combined concepts and attributes. In the fifth (training) block, participants had to categorize attributes again, but this time the attributes switched positions on the screen (e.g., when the positive attribute was presented on the left side of the screen in block 2, it was now positioned on the right side of the screen). Critical blocks 6 and 7 once again presented participants with the combined concepts and attributes, but this time the position of the attributes was switched relative to the position in critical blocks 3 and 4 (i.e., if "Dutch" + negative was presented on the left side in block 3 and 4, it was now presented on the right side).

For both the WIAT and PIAT, participants were issued one of four versions that varied on the following randomized factors: the starting position of the concept (i.e., left side or right side of the screen), and whether the concept is expected to be congruent or incongruent with the outgroup negativity bias. See Table S3 for more details.

In the word-IAT, words were used to indicate the superordinate categories (concepts and attributes), and pictures for the to-be categorized stimuli. For the superordinate categories, the concepts were written in a black font as "Nederlands"

(*Dutch*) or "Marokkaans" (*Moroccan*). The attributes were written in a green font as "positief" *(positive)* and "negatief" (*negative*). In the critical blocks where combined categories were presented, the concept and attribute words were written on top of each other, separated with an "of" (*or*). Their order (top or bottom) was randomized. The to-be categorized stimuli were always presented in the lower part of the screen in the center, just like in the PIAT in Experiment 1, and consisted of the same images as in Experiment 1. The PIAT in Experiment 2 was similar to the one in Experiment 1, with the difference being that there were now two more critical blocks and answers were given via pressing "E" and "I", rather than touching the screen.

Symbolic Racism 2000 Scale

The Symbolic Racism 2000 Scale (henceforth: SRS) was created to assess explicit inter-ethnic biases via a series of eight questions (Henry & Sears, 2002). The SRS specifically looks at a modern variant of discrimination in the form symbolic racism, or the belief that discrimination based on ethnicity is no longer impacting people of non-Dutch descent's chances to thrive and that continuing disadvantages are attributable to their lack of responsibility for their own lives (Henry & Sears, 2002). The SRS issues participants with questions regarding work ethic and responsibility of outcomes, excessive demands, denial of continuing discrimination, and undeserved advantages. Participants answer most questions on a 4-point Likert scale ranging from "strongly disagree" to "strongly agree", but questions 3, 4 and 5 involve different types of answers. The questions in the SRS are specifically attuned to the cultural history of people of color residing in the United States. To make the SRS applicable to our study group, we translated the English questions to Dutch, and replaced the words "African-American", "United States", "Irish, Italian, Jewish, and many other minorities" to "Marokkaans" ("Moroccan"), "Nederland" (The Netherlands), "Mensen met een Surinaamse of Poolse afkomst, of andere minderheden" ("People of Surinamese or Polish descent, and other minorities"), respectively.

Stimuli

The stimuli were the same as the ones used in the PIAT in Experiment 1. However, whereas in Experiment 1 there were different versions of the PIAT using different exemplars for the attribute categories, in Experiment 2 we used only one exemplar for the positive and negative attributes (i.e., the rabid dog and the seal pup from the IAPS) to ensure that we had enough participants per versions of the tasks. The PIAT and WIAT consisted of four versions, which each differed in a) whether incongruent

trials were presented first or second, and b) the location of the stimuli on the screen (e.g., faces of men of Moroccan or Dutch descent, or positive/negative images and text on the left/right side of the screen, see Table S3 in supplements for an overview of all task versions). All images were presented in color, and all images were presented a maximum of two times during the two critical blocks.

Equipment

Experiment 2 was conducted online through Qualtrics and by using Iatgen, a pre-programmed survey-software implicit association test (Carpenter et al., 2019). As the original survey software uses words only, we adapted it to work with pictures. Participants using mobile devices such as tablets and smartphones were not allowed to participate in the study.

Procedure

Participants who signed up to take part in the study were sent a link to the tasks. Via Qualtrics, participants first received brief information about the goal of the study, namely to compare two types of categorization tasks containing faces and scenes. They were also told that they would receive more information about the study after completing the experiment. If participants were still interested in participating, they signed a digital consent form to allow us to use their data. Next, participants were issued questions about their age, gender, handedness, native language, and their own and their parents' nationality. The study would be terminated with a custom message if the participants were below of the age of 18, or if they or their parents did not have the Dutch nationality. If participants passed the screening, they were notified the first categorization task would start when they continued on to the next screen. To keep the online experiment as similar as possible to the PIAT in experiment 1, this was all the information that was given.

Every participant took part in both the WIAT and the PIAT. The order of the IATs was counterbalanced, and in-between each IAT the participant completed the SRS. Between each of the tasks, participants could decide to take a break for as long as they wanted. At the end of the experiment, participants were asked questions about their prior experience with people of Moroccan descent, i.e., whether they knew anyone of Moroccan descent and if they did, how well they knew this person or these persons. This information was not used in subsequent analyses, as it was part of a different research project. Participants were then given a full debriefing on the goals of the study, and were thanked for their participation.

Analyses

D-score calculations and analyses were performed in R, using the IATScores package (Richetin et al., 2015). As per the suggestion of Richetin et al. (2015), we did not distinguish between the first 20 trials ("practice" trials) in the critical block and the 40 following trials. The minimum performance criteria were an error rate below 40% for the critical blocks, and reaction times (RTs) higher than 400 ms and lower than 10,000 ms (Nosek et al., 2014). RTs below or above these criteria were discarded. Based on these criteria, we discarded 4 trials with an RT > 10,000, and 23 trials with an RT < 400 (divided relatively equally across 13 subjects) within the PIAT sample. For the WIAT sample, we removed 6 trials (divided over 4 individuals) based on the 10,000 ms cutoff, and 124 trials based on the 400 ms cutoff (divided over 16 individuals, of which one individual had 20 trials meeting this criterion, and another individual 80 trials). No participants were removed based on the 40% error rate cutoff in the PIAT, and one participant in the WIAT. This method is slightly different from Experiment 1, where only RTs above 10,000 ms were discarded and remaining RTs were 10% winsorized. While this is the most robust approach for treating IAT data (Richetin et al., 2015), this was not an option due to an error in data collection. During data collection, trials with where a wrong categorization was made were flagged by the software, but due to an error on our part, RTs for these trials were not saved and thus erroneous trials could not be included in the analysis. Instead, we discarded RTs lower than 400 ms (as per the original scoring method by Greenwald et al., 2003).

We used one sample t-tests to test whether D-scores significantly differed from zero. Furthermore, to assess internal validity, we also fit linear models using *Congruency* (incongruent first vs. congruent first) and *Location* (whether faces of men of Moroccan or Dutch descent were presented left or right) as fixed effects (sum-to-zero coded) in two separate analyses (PIAT and WIAT), with the intercepts reflecting the average D-score. Note that instead of *Location*, we used *Task Version* as fixed effect in Experiment 1. This was done because the versions in Experiment 1 differed on multiple fronts (i.e., stimuli and location on the screen), whereas in Experiment 2, versions of the tasks only differed in the location of the stimuli on the screen. Next, we correlated D-scores on the PIAT with the WIAT in order to assess test-retest reliability. In addition to comparing the IATs to each other, they were also compared with the explicit bias measure (SRS) in order to assess discriminant validity. The data of the SRS were first converted to a continuous 0-1 scale as described by Henry & Sears (2002), and subsequently used for a correlation calculation. Finally, we also investigated whether providing the SRS in-between the two tasks (and thus before one of the two

tasks) affects subsequent D-scores on the final task (e.g., due to a priming effect). For this, we first calculated a difference score between performance on the tests by subtracting the D-score from the second task from the first. We then fitted a linear model using *First Task* (PIAT first or WIAT first) a sum-to-zero coded fixed effect and *Difference Score* as the dependent variable. Furthermore, see supplemental Figure S5 for our sensitivity power analysis results.

Results

Implicit associations

For the PIAT, we found a significantly positive D-score average of .18 (*95% CI* [.07, .28], $t(139) = 3.34$, $p = .001$, *Cohen's d* = .28), meaning that participants associated faces of men of Moroccan descent with negativity, and faces of men of Dutch descent with positivity (Figure 4). In the linear model controlling for *Congruency* and *Location*, this effect remained present (D-score: .17, *95% CI* [.07, .27], $t(137) = 3.36$, $p = .001$). We also found a significant effect of *Congruency* on D-scores ($F(1, 137) = 7.03$, $p = .009$),

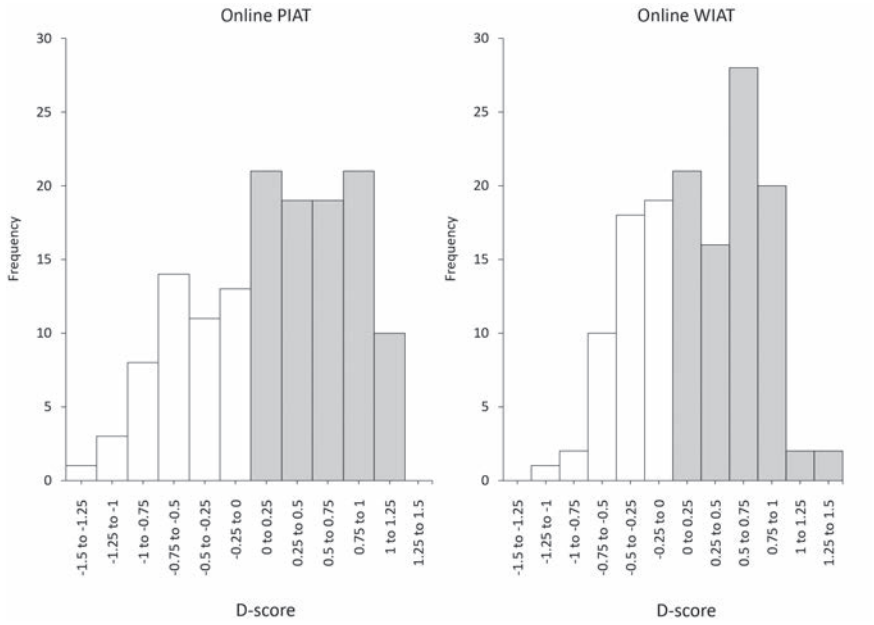

Figure 4. D-score distribution for the online PIAT (left) and online WIAT (right). Positive values represent stronger associations between faces of men of Dutch descent and positive scenes, and faces of men of Moroccan descent and negative scenes.

with individuals receiving congruent trials first having a higher D-score on average ($M = .31$, $SD = .64$) than individuals receiving incongruent trials first ($M = .04$, $SD = .58$, supplemental Figure S6 and Table 3). Finally, we did not find an effect of *Location* on D-score averages ($p = .194$, Table 3).

Results of the WIAT indicate a significantly positive D-score average of .22 (*95% CI* [.13, .31], $t(138) = 4.91$, $p < .001$, *Cohen's d* = .42), similar to what the PIAT showed (Figure 4). In the linear model this finding held up (D-score: .22, *95% CI* [.13, .31], $t(136) = 4.88$, $p < .001$). However, there was no *Congruency* effect ($p = .269$), nor an effect of *Location* ($p = .409$, supplemental Figure S6, Table 3).

Finally, there was a strong positive correlation between individuals' scores on the PIAT and WIAT ($N = 139$, $r = .69$, $p < .001$, Figure 5), indicating that participants who had an implicit bias in one of the tasks showed a similar bias in the other task.

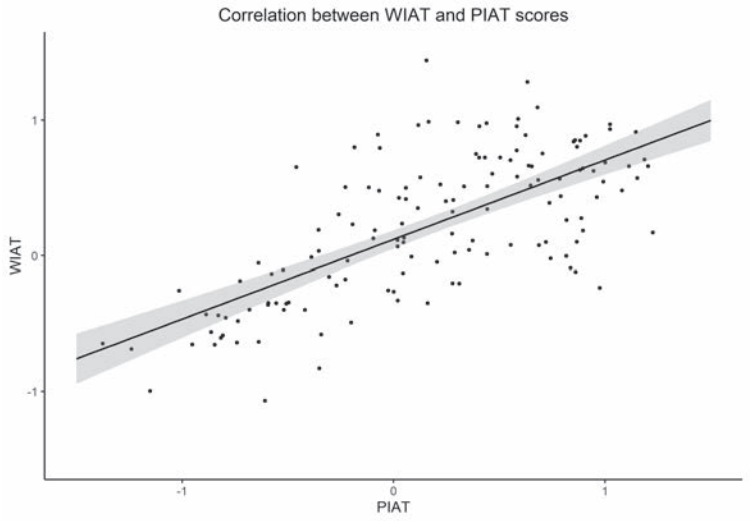

Figure 5. Scatter plot showing the correlation between WIAT and PIAT D-scores ($r = .69$, $p < .001$).

Correlation between the WIAT, PIAT and explicit measures

All 144 participants completed the Symbolic 2000 Racism Scale. On average, individuals had an SRS score of .33 ($SD = .13$, with scores ranging between .04-.64), meaning that explicit symbolic racism among our participants was low (Henry & Sears, 2002). For both the WIAT and the PIAT, we did not find a significant correlation between implicit biases (D-scores) and explicit symbolic racism score (WIAT: $r = -.02$, $p = .801$; PIAT: $r = -.01$, $p = .934$). To test whether our results indicated evidence for the null-

hypothesis, we performed an additional correlation analysis using Bayesian statistics (using the R-packages BayesFactor (Rouder et al., 2009) and bayestestR (Makowski et al., 2019b). We found a $BF_{01} = 5.1$ for the correlation between PIAT and SRS scores, and a $BF_{01} = 4{,}95$ for the correlation between WIAT and SRS scores, meaning that in both cases the data are around 5 times more likely under the null hypothesis that there is no correlation between the variables. The results therefore indicate moderate evidence for the null hypothesis (i.e., there is no correlation between the measures (Lee & Wagenmakers, 2013). As such, the implicit biases reflected in the D-scores in both versions of the IAT do not seem to correlate with explicit inter-ethnic biases in our participant pool.

Finally, we also assessed whether performing the SRS in-between tasks affected D-scores on the final IAT. We found that the difference score did not significantly differ from zero ($t(137) = -1.14$, $p = .255$), meaning that D-scores in the second task were not significantly higher or lower than the D-scores in the first task. Furthermore, we found no significant effect of *First Task* ($t(137) = -1.27$, $p = .206$) on the difference score, showing that the difference score was not affected by whether a participant first started with the PIAT or the WIAT. In short, performing the SRS in the middle of two IATs did not significantly impact D-scores on the last IAT, and this was regardless of whether participants started with a PIAT or WIAT.

Table 3. Model results for the online PIAT

Task		D-score				
PIAT	*Predictors*	*Estimates*	*SE*	*95% CI*	*t*	*p*
	(Intercept)	.17	.05	.07 – .27	3.36	**.001**
	Congruency (Congruent first)	-.14	.05	-.24 – -.03	-2.63	**.009**
	Location ("Moroccan descent" Left/ "Dutch descent" right)	.07	.05	-.03 – .17	1.31	.194
WIAT	(Intercept)	.22	.04	.13 – .31	4.88	**<.001**
	Congruency (Congruent first)	-.05	.04	-.14 – .04	-1.11	.269
	Location ("Moroccan descent" Left/ "Dutch descent" right)	-.04	.04	-.13 – .05	-0.83	.409

Conclusion

Participants performed similarly on the PIAT and WIAT: in both IATs, participants appeared to have a an implicit bias in the predicted direction. Although the order in which participants completed the critical blocks significantly impacted the D-scores

in the PIAT (i.e., participants with congruent trials first had a higher D-score average than participants with incongruent trials first), this effect was not present in the WIAT. The PIAT and WIAT results were significantly correlated and showed good (within-participant) test-retest reliability. Finally, implicit biases measured with the IATS did not correlate with explicit measures of inter-ethnic bias measured through the SRS, nor did the SRS impact performance on the IAT that followed it. The results extend our findings from Experiment 1, showing that the PIAT can tap into the same implicit biases as the more commonly used WIAT.

Discussion

The aim of this study was to design and validate a non-verbal, intuitive pictorial IAT (PIAT). Experiment 1 shows that the PIAT can tap into implicit inter-ethnic attitudes in a large group of participants including adults and children, and can do so reliably using different stimuli. Furthermore, the PIAT can do this outside of a lab setting and on a representative subject population involving participants of different ages (as opposed to only university students). In Experiment 2 using a within-subjects design, the performance of the PIAT was comparable to a more typical word IAT that has been rigorously tested in the last two decades (Dunham et al., 2006; Greenwald et al., 1998, 2003; Kurdi et al., 2019; Nosek et al., 2002a, 2013; Oswald et al., 2015). As such, the PIAT could be standardized tool that enables future studies to make direct comparisons across different cultures, age groups, and potentially also between species.

Internal validity

Although we counterbalanced the order of the presentation of critical blocks, participants who received incongruent trials first and congruent trials second in the PIAT in Experiment 1 showed higher D-scores on average than participants who received a reversed order. Interestingly, this effect was reversed in the PIAT of Experiment 2, and absent in the WIAT. IAT order effects are well documented, and may impact the magnitude of the found IAT effects (Greenwald, 2001). An explanation for their existence is a cost of switching tasks in the two critical blocks, namely in the form of increased reaction time latencies and error rates (Mierke, 2001; Mierke & Klauer, 2003). Furthermore, these task switching costs may remain for quite some time after switching, as switching requires the activation of the appropriate action and suppressing the previous, competing one. Nevertheless, we find diametrically

opposed results in the order effects, which suggests that order effects likely occurred due to noise or random differences between the groups of participants. Our data therefore do not clearly support an effect of block sequence on IAT scores. Order effects remain a topic of debate in the IAT literature, and thus deserve more attention in future studies.

Psychometrics (internal consistency)

The PIAT shows an acceptable internal consistency in children, and good internal consistency in adults. The result of the child PIAT is somewhat lower than internal consistencies revealed in a comparable PIAT used in children (e.g., $\alpha = 86.5$ on average across two PIATs (Cvencek et al., 2011)), which could be explained by the more noisy setting in which the PIAT was distributed. Generally, our results are in line with the limited amount of studies using image-based IATS that report internal consistency values (Brand et al., 2014; Palfai et al., 2016; Slabbinck et al., 2011). Importantly, despite the less-controlled circumstances in which we conducted the experiment, the PIAT reveals ethnicity-based implicit associations consistent with previous findings (Dunham et al., 2008; Greenwald et al., 1998).

Choice of stimulus material

For all IATs, we created several different versions of the task that differed on the stimuli that represented the attribute dimensions (in Experiment 1: PIAT Version 1, 2 and 3), or differed in the order of critical block presentation and the location of the concepts and attributes on the screen (All IATS in Experiment 1 and 2). We found that children showed a lower D-score average in version 1 of the PIAT in Experiment 1, and that the spread in D-scores was higher in this version than in the other versions. This effect was, however, not present in the adult PIAT nor in the online PIAT and WIAT. It is important that the differences in the attribute dimensions are clear (i.e., clearly negative and positive) as the IAT effect relies heavily on responses that do not require a lot of deliberation (Lane et al., 2007), which is why we used different types of stimuli that were rated by children in Experiment 1 on valence and or whether they were low or high in arousal. We found that the ratings for the different stimuli were very close to each other (see Table 1c in supplements), thus it is unclear what it is about the stimuli used in task version 1 that resulted in lower D-score averages compared to the other versions. The IAT represents a family of instruments where differences in e.g., choice of stimuli can result in entirely different results, even when the versions were built with the aim to measure the same underlying construct (Foroni & Bel-Bahar,

2009). As such, it remains crucial to choose stimuli that reflect the same superordinate category that one is interested in as much as possible (Lane et al., 2007), but also to include several different stimuli (Nosek et al., 2005) as we aimed to do in our study.

Discriminant validity

Both the PIAT and WIAT do not correlate with the explicit measure of inter-ethnic bias (the symbolic racism scale, or SRS), which suggests that IATs tap into a different kind of cognitive constructs than explicit biases. Indeed, this is the reason why IATs exist in the first place, namely to show that views and attitudes are partially driven by unconscious mechanisms (Greenwald et al., 2002). At the same time, we did not find high antipathy scores against individuals of Moroccan descent, which could for instance reflect that a) our participants indeed do not consciously view individuals of Moroccan descent as more negatively than individuals of Dutch descent, or b) participants answer in a socially desirable way, which can be true especially in case of a highly sensitive topic such as ethnic prejudice (Fazio & Olson, 2003). In the current study we cannot directly dissociate between these explanations, but a meta-analysis on the correlation between implicit and explicit measures of attitudes showed that correlations between implicit and explicit measures may be low when participants make their judgments deliberately or spontaneously. For instance, it takes more cognitive effort when asked to reflect on your evaluations of individuals with differing backgrounds than whether you are asked about more mundane things such as attitudes towards fruits and candies; in the latter case, correlations between explicit an implicit measures are higher (Hofmann et al., 2005) .

In general, IATs that find ethnicity biases do indeed report negative correlations between IAT effects and explicit biases, and our results are in line with these findings. Furthermore, the notion that the PIAT indeed uncovers implicit biases is supported by the fact that the Symbolic 2000 Racism Scale questionnaire, which primes participants to think more deeply about their inter-ethnic biases, did not seem to affect subsequent IAT effects. For further discussions on the correlations between explicitly and implicitly assessed attitudes we refer to the meta-analysis by Hofmann et al. (2005).

PIAT performance compared to the WIAT

In Experiment 2 we show that the PIAT performs similarly to a word-IAT, and that the test-retest reliability of the IAT measures was good. This is interesting considering the persistent debate on the relatively large range of test-retest reliability scores of the IAT (Lane et al., 2007). The benefit of a PIAT over IATs that use words or spoken

language are that it is applicable to a wider variety of populations, and that the same test (i.e., without having to translate words) can be used even in populations that are very different (e.g., because of culture, language, cognitive ability), thus making direct comparisons possible. The stimuli we used for the (adult) PIAT were selected from the cross-culturally validated International Affective Picture System (Lang et al., 2007), and the IAPS is one example of what researchers can use to study e.g., cultural differences in implicit attitudes. At the same time, while the PIAT has not been validated for non-human animals, it could potentially be a useful tool to study implicit attitudes in for instance great apes, as they are highly capable of extracting emotionally relevant information from scenes and can be trained on the use of a touchscreen (Altschul et al., 2017; Kret et al., 2016; Perdue et al., 2012). For animals, more appropriate positive and negative images should then be selected (e.g., a favorite food item or an item that holds a negative association).

Conclusion

With the aim of validating a non-verbal PIAT, we found that it can be used to measure implicit biases reliably, and similarly to a standard verbal IAT. As such, we believe it can provide a practical way to study implicit associations in a wide variety of individuals, and conceivably in non-verbal populations. Pictorial adaptations to the IAT have the potential to answer important questions related to the ontogeny and evolutionary development of implicit attitudes, and to directly compare different groups of individuals on their implicit associations. Intergroup conflict in humans is still ubiquitous, and the discussions about the foundations of implicit associations are still ongoing. We therefore deem it crucial to find novel ways to probe these implicit attitudes and make within- and between-species comparisons possible, which we think is the most important role that pictorial adaptations to the IAT can fulfill. By validating the PIAT as a tool, our study sets a first step into that direction, but future studies should look into optimization of the task by testing different kinds of attitudes and using multiple different category exemplars. Ultimately, we hope the PIAT can be added to the steadily growing list of cognitive tasks that can be used in comparative research.