Available online at www.sciencedirect.com
ScienceDirect
Journal of Experimental Social Psychology 43 (2007) 972–978
Journal of Experimental Social Psychology
www.elsevier.com/locate/jesp

Novel attitudes can be faked on the Implicit Association Test ☆

Jan De Houwer [a,*], Tom Beckers [b], Agnes Moors [c]

[a] Department of Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium
[b] Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium
[c] Department of Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium

Received 12 June 2006; revised 13 September 2006
Available online 22 November 2006
Communicated by Hart Blanton

Abstract

We asked participants to imagine that a researcher would provide them with positive or negative information about fictitious social groups. Half of the participants were asked to act in such a way that they would conform to the expectations of the researcher. The other participants were asked to behave in the manner opposite to what the researcher expected. Participants then completed an IAT designed to measure the newly formed attitudes toward the fictitious social groups. The direction of the IAT effect depended on the faking instructions. The results call for caution when using the IAT to study the development of implicit attitudes.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Attitude formation; Implicit attitudes; IAT

Since the influential work of researchers such as Zajonc (1980), Fazio (1990), and Greenwald and Banaji (1995), it is generally accepted that implicit attitudes (i.e., automatic evaluative responses) have an important impact on behavior. In order to understand, predict, and control behavior, researchers must therefore be able to measure implicit attitudes and know how these attitudes are formed and can be changed. A number of tasks have been developed that can be used to measure implicit attitudes (see Fazio & Olson, 2003, for a review), the most popular of which is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). Most often, these tasks have been used to measure existing attitudes in an attempt to understand and predict behavior. Recently, however, the IAT and other implicit measures have also been used to study the development and modification of implicit attitudes. This important research hinges on the assumption that existing implicit measures are valid indicators of newly developed implicit attitudes. The aim of our research is to examine this assumption, more specifically with regard to the IAT. Because the focus of our research is on the IAT as a measure of novel implicit attitudes, we will now summarize previous studies in which the IAT has been used in this manner.

In the studies of Olson and Fazio (2001, Experiment 2), participants were first exposed to a learning phase during which the names and drawings of various unknown pokemon characters were presented together with positive, negative, and neutral pictures and words. One pokemon character was paired consistently with positive words and pictures, whereas a second pokemon was always presented together with negative words and pictures. During a subsequent test phase, participants performed an IAT during which stimuli appeared on the screen that belonged to one of four categories: (1) positive words, (2) negative words, (3) the picture or name of the pokemon that was paired with positive stimuli, and (4) the picture or name of the pokemon that was paired with negative stimuli. Participants were asked to press one of two keys depending on the

☆ Jan De Houwer and Agnes Moors, Ghent University, Belgium. Tom Beckers, University of Leuven, Belgium. Tom Beckers and Agnes Moors are post-doctoral researchers of the Research Foundation—Flanders (FWO-Vlaanderen).
* Corresponding author. Fax: +0032 9 264 64 89.
E-mail address: Jan.DeHouwer@UGent.be (J. De Houwer).

0022-1031/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2006.10.007

category to which the presented stimulus belonged. There were two tasks. In the compatible task, the pokemon that was previously paired with positive stimuli had to be classified in the same manner as positive words whereas the pokemon that was paired with negative stimuli was assigned to the same key as negative words. In the incompatible task, the assignments were reversed (press the first key for positive words and the pokemon that was previously paired with negative stimuli; press the second key for negative words and the pokemon that was previously paired with positive stimuli). Results showed that participants were faster in the compatible than in the incompatible task. This IAT effect suggests that implicit attitudes as measured by the IAT can be formed by pairing previously unknown stimuli (e.g., pokemons) with already liked or disliked stimuli (e.g., positive or negative pictures and words) (see De Houwer, Thomas, & Baeyens, 2001, for a review of related findings).

Whereas Olson and Fazio (2001, 2002) did not inform the participants that there was a relation between the pokemons and the positive and negative stimuli during the learning phase, in other studies, the nature of the relation between the neutral and valenced stimuli was made explicit. In two studies, Mitchell, Anderson, and Lovibond (2003) first told participants that they would see new words together with their meaning in English. During a subsequent learning phase, two new words were repeatedly presented together with a positive word and two other new words were repeatedly paired with a negative word. Finally, an IAT was used to assess the implicit attitude toward the new words. In both studies, IAT effects revealed more positive attitudes toward the new words that were said to have a positive meaning than toward new words that were said to have a negative meaning.

Gregg, Banaji, and Seibt (2006) informed their participants that they would receive information about the characteristics of two unknown social groups called *Niffites* and *Luupites*. All *Niffites* had a name that contained a double consonant and ended with *nif* (e.g., Cellanif). All *Luupites* had a name that contained a double vowel and ended with *lup* (e.g., Maasolup). A group was described either in a positive (i.e., very good, peaceful, civilized, benevolent, and law-abiding) or negative manner (i.e., its members were very bad, violent, savage, malicious, and lawless). A subsequent IAT test showed that participants had a more positive attitude toward the group that was described in a positive manner than toward the group that was described in a negative manner. Interestingly, this effect was found even for other fictitious socials groups that were said to be equivalent to *Niffites* or *Luupites*. This led Gregg et al. to the conclusion that implicit attitudes can be formed on the basis of mere supposition.

In line with the findings of Gregg et al., De Houwer (2006a) recently found that merely instructing participants that nonwords (Experiment 1) or names of fictitious social groups (i.e., *Niffites* and *Luupites*; Experiment 2) would be consistently followed by either positive or negative pictures was sufficient to change the attitudes toward these objects as measured by the IAT. However, rather than concluding that implicit attitudes can be formed on the basis of instructions, De Houwer also raised the possibility that the observed IAT effects were due to demand compliance. This argument might seem implausible given existing research suggesting that IAT effects are difficult to fake, provided that it is the first time that one completes an IAT (e.g., Fiedler & Bluemke, 2005; Kim, 2003; Steffens, 2004). However, all existing studies on faking attitudes in the IAT have involved well-known attitude objects for which participants probably already had well-established attitudes (e.g., black persons, homosexuals). It therefore remains to be seen whether attitudes toward novel attitude objects are also difficult to fake on the IAT.

In order to examine this issue further, we conducted a new study in which participants were first informed about the design and results of the study by Gregg et al. (2006). We then pointed out that these results could have been due to demand compliance, that is, to the fact that the participants of Gregg et al. deliberately acted as if their attitudes were in line with the expectations of the researchers. As a test of this hypothesis, the participants in our study were asked to act as if they had attitudes towards fictitious social groups that were consistent with the information given about these groups. They then performed an IAT designed to measure attitudes toward the social groups. If we observe that the IAT reveals attitudes toward the social groups consistent with the information provided about the social groups, this would suggest that attitudes toward novel attitude objects can be faked. One could argue that such a result could also arise if implicit attitudes are still formed on the basis of novel information even when the participants are told that the information is false. For instance, Gawronski and Bodenhausen (2006) recently argued that the negation of propositional information influences only explicit and not implicit attitudes.[1] Following a suggestion of an anonymous reviewer, we therefore also included a second condition in which participants were given the same information as in the other condition but were asked to behave in a manner opposite to the information provided about the groups. If the IAT effect in the latter (inconsistent faking) condition is significantly different from the effect in the former (consistent faking) condition, this would show that IAT performance is determined by the faking instructions rather than by the information about the groups.

[1] Note, however, that based on the model of Gawronski and Bodenhausen, it is unlikely that verbal instructions of the kind we used would result in the formation of implicit attitudes. According to their model, new implicit attitudes are formed gradually over many experiences rather than as the result of new propositional knowledge that is presented in instructions.

Method

Participants

Seventy-two students from various faculties at Ghent University and 62 first year psychology students at the University of Leuven took part in this study. Students at Ghent University were paid 5 euros, those at the University of Leuven participated for course credit. All participants were native Dutch speakers.

Materials

The social groups were called Niffites and Luupites. The Niffite names were Cellanif, Eskannif, Lebbunif, and Zallunif, and the Luupite names were Meesolup, Naanolup, Omeelup, and Wenaalup. These were the same labels and names as those used by Gregg et al. (2006). During the IAT, the eight names as well as four positive words (the Dutch words for FLOWERS, HUMOUR, HOLIDAY, and SPRING) and four negative words (the Dutch words for SADNESS, STENCH, CANCER, and DEATH) were presented in uppercase letters using the font Arial Black with font size 36. The experiment was controlled by an Inquisit 1.32 program that was implemented on a Dell PC with a Pentium 4 processor and a 15 in. screen. Participants responded by pressing the Q or M key on an AZERTY keyboard.

Procedure

Participants took part one by one. They first read instructions which informed them about the results of Gregg et al. (2006) and were then told that these results might have been due to demand effects. We asked our participants to imagine that a researcher gave them information about fictitious social groups. Half were told that Niffites were good and Luupites were bad. The other participants received the opposite information. Within each of these conditions, half of the participants were asked to act as if they had attitudes that were consistent with the provided information (consistent faking condition) and half were instructed to act as if their attitudes were inconsistent with the information about the social groups (inconsistent faking condition). A full transcript of the instructions can be found in the Appendix.

After reading the instructions, participants completed a rating task and an IAT. The order of the two tasks was counterbalanced over participants but did not influence the IAT results. In the instructions for the rating task, participants were asked to indicate for each Luupite and Niffite name their feelings about that Luupite or Niffite. They could do this by clicking on a number from 1 to 9 where 1 stands for "very negative" and 9 stands for "very positive." The eight names were then presented in a random order that was determined separately for each participant. On each rating trial, the name appeared in the centre of the screen. A 9-point rating scale of 25 cm long was presented beneath the name. The next trial started 1000 ms after the participants had clicked on a number. Given that the rating data were of secondary interest, we will present them in a footnote only.[2]

At the start of the IAT, participants were informed that words would appear one by one on the screen and that word meaning would always be important. A word could belong to one of four categories: (1) positive words, (2) negative words, (3) names of Luupites, and (4) names of Niffites. The task of the participant was to press a left (Q) or right (M) key depending on the category of the presented word. Participants were told that the assignment of categories to responses would vary from phase to phase. The names of the categories assigned to the left key in a particular phase would always be presented in the left top corner of the screen and the names of the categories assigned to the right key would always appear in the right top corner. Finally, participants were urged to respond as quickly as possible without making too many errors.

The IAT always started with a practice block of 32 trials during which the 8 names (4 Luupites and 4 Niffites) were each presented 4 times. Next came a practice block of 16 trials during which the 8 words (4 positive and 4 negative) were each presented twice. The practice blocks were followed by 2 test blocks of 32 trials. During each test block, each of the 8 names and 8 words was presented on 2 trials. The practice block with names was then repeated but the response assignments were reversed. Finally, the two test blocks were repeated with the reversed response assignment for the names. The order of the trials was determined randomly for each block and each participant separately.

Each block was preceded by instructions about which type of stimuli would appear in the next block and about which categories were assigned to which keys. During a block, the category labels (POSITIVE, NEGATIVE, LUUPITES, and NIFFITES) were presented in the corresponding top corner of the screen. Which categories were assigned to which response in which phase was counterbalanced over participants. Half of the participants always had to press the left key for positive and the right key for negative words whereas the other participants pressed the left key for negative words and the right key for positive words. Orthogonal to this manipulation, half of the participants started by pressing the left key for Luupites and the right key for Niffites, whereas the other half started with the opposite assignment. By counterbalancing these factors in this manner, half of the participants started with the

[2] For each participant, we calculated a rating difference score by subtracting the mean rating for the names that belonged to the so-called bad social group from the mean rating of the names that belonged to the so-called good social group. One sample t-tests showed that the rating score was positive in the consistent faking condition, $M = 4.91$, $SD = 2.99$, $t(66) = 13.47$, $p < .001$, $d = 1.64$, showing attitudes in line with the provided information. The rating score in the inconsistent faking condition was negative, $M = -4.00$, $SD = 4.05$, $t(67) = 13.47$, $p < .001$, $d = 0.99$, which indicates attitudes inconsistent with the provided information.

Table 1
Mean untransformed reaction times in ms as a function of IAT task and condition

Condition	IAT task								
	Compatible			Incompatible			IAT effect		
	M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
Consistent faking	759	150	518/1256	912	227	574/1642	159	165	−66/637
Inconsistent faking	860	212	600/1586	809	166	554/1268	−51	169	−540/264

compatible task (i.e., the social group that was said to be positive was assigned to the same key as positive words whereas the social group that was said to be negative was assigned to the same key as negative words). These participants afterwards performed the incompatible task (i.e., the so-called positive social group was assigned to the same key as negative words whereas the so-called negative social group was assigned to the same key as positive words). The other half of the participants first completed the incompatible task and afterwards the compatible task. The manipulation of the order of the IAT tasks was orthogonal to the other manipulations (i.e., type of information; type of faking instructions; order of tasks).

Before each set of two test blocks, participants were reminded about the expectations of the researcher (e.g., "REMINDER: Do not forget that the researcher expects that you act as if you like NIFFITES and dislike LUUPITES.") and asked to act as if they felt those feelings (consistent faking condition) or to act as if they felt the opposite feeling (inconsistent faking condition). These reminders were added to make sure that the participants would not forget the information about the social groups and their task.

On each trial during the IAT, a word was presented in the centre of the screen until the participant pressed one of the two valid keys (Q or M). If the response was correct, the word disappeared and the next word was presented 400 ms later. If the response was incorrect, the word was replaced by a red X for 400 ms. The next word appeared 400 ms after the red X was removed from the screen.

Results

One participant in the inconsistent faking condition made more than 40% errors in both IAT tasks, probably because of failure to understand the instructions. Six other participants in the inconsistent faking condition made more than 40% errors only in the task where the social group which was said to be good was assigned to the same key as positive words. A closer inspection of the data showed that they almost consistently responded to the names of the social group that was said to be good by pressing the key for negative words and to the names of the other social group by pressing the key for positive words. Hence, these participants apparently responded to the instruction to act in opposition to the provided information by explicitly evaluating the names in a way opposite to the information provided about the groups. The data of these seven participants were excluded from the analyses, leaving 67 participants in the consistent faking condition and 61 participants in the inconsistent faking condition.

For each participant, we calculated the D600 IAT difference score in line with the recommendations of Greenwald, Nosek, and Banaji (2003).[3] The score was calculated in such a way that a positive value indicated faster performance in the compatible task (the social group that was said to be good was assigned to the same key as positive words and the other group to the same key as negative words) than in the incompatible task (the good social group was assigned to the same key as negative words and the bad social group to the same key as positive words). For ease of interpretation, we report the mean untransformed reaction times and difference scores in Table 1.

An ANOVA with faking instructions (consistent or inconsistent) as a between-subjects variable and the D600 IAT difference score as the dependent variable revealed a main effect of faking instructions, $F(1, 126) = 60.17$, $p < .001$. One sample t-tests revealed a large positive IAT effect in the consistent faking condition, $t(66) = 9.14$, $p < .001$, $d = 1.12$, which indicates attitudes in line with the information provided about the social groups. The distribution of the IAT effects in this condition was normal with 55 of the 67 participants showing an effect in the expected direction. On average, participants gave an incorrect response on 4.95% of all trials (range 0–12.50%). The IAT effect in the inconsistent faking condition was small but significant and in the negative direction, $t(60) = -2.54$, $p = .01$, $d = 0.33$, indicating attitudes opposite to the information provided about the groups. Again the distribution of the IAT effects in this condition was normal with 34 of the 61 participants showing an effect in the expected direction (i.e., negative IAT scores). The mean percentage of incorrect responses was 6.26% (range 0–27.50%).

To compare the size of the IAT effect in the consistent and inconsistent faking condition, we multiplied the IAT effect score in the inconsistent faking condition with −1 so that in both conditions, a positive score indicates successful faking. When we repeated the ANOVA on these transformed IAT scores, we again found a main effect of faking instructions, $F(1,124) = 13.85$, $p < .001$, other Fs < 1. This main effect shows that in absolute terms, the IAT effect was

[3] The same conclusions were reached when IAT effects were calculated on the basis of log-transformed reaction times as recommended by Greenwald et al. (1998).

stronger in the consistent faking condition than in the inconsistent faking condition.

Discussion

Our results suggest that participants can fake having newly formed attitudes as measured by the IAT. The IAT revealed attitudes in line with the information provided about fictitious social groups when participants were asked to fake attitudes that were consistent with this information. The fact that this IAT effect was due to the faking instructions rather than to the provided information was confirmed by the observation that a different IAT effect was obtained when participants were asked to fake attitudes that were inconsistent with the provided information. In this case, the IAT effect suggested attitudes opposite to the information that was provided. These findings not only increase our understanding of the factors that determine the outcome of the IAT, they also have implications for research in which the IAT is used as a measure of novel (implicit) attitudes. Because IAT effects for novel attitude objects can be faked, one has to be careful when interpreting an IAT effect as evidence for a novel attitude. For instance, using the IAT as a measure of novel implicit attitudes, Gregg et al. (2006) concluded that merely supposing that novel social groups have certain evaluative characteristics is sufficient for the formation of implicit attitudes toward these groups. Our results, however, suggest that the results of Gregg et al. could also have been due to demand effects. This does not imply that the IAT cannot detect novel implicit attitudes. In fact, the studies by Olson and Fazio (2001) suggest that the IAT can detect such attitudes. In those studies, participants were not aware of the fact that novel attitude objects were paired with positive or negative stimuli. Hence, it was impossible for them to know which attitudes they should fake. However, when participants are aware of the way in which novel attitude objects are linked with positive or negative information, it is important to realize that one cannot avoid demand effects by using the IAT as a measure of the novel attitudes.

Our results also illustrate the more general point that one cannot simply assume that implicit measures of attitudes are implicit in all possible meanings of the word. As De Houwer (2006b; De Houwer and Moors, in press) recently pointed out, it needs to be examined empirically whether and in which sense a measure is implicit. One of the ways in which a measure can be implicit is that novel attitudes cannot be faked. Our findings suggest that an IAT measure of novel attitudes is not implicit in that sense of the word. Whether the same conclusion holds for other implicit measures is a matter for future research. But it is imperative that this is examined empirically before any claims are made about the formation of attitudes on the basis of these implicit measures.

The present data provide particularly compelling evidence for the hypothesis that attitudes can be faked on the IAT. Until now, evidence for faking in the IAT was found only when participants performed the IAT for a second time immediately after completing the same IAT for the first time (Fiedler & Bluemke, 2005; also see Steffens, 2004). The present study is the first to demonstrate successful faking in the IAT when participants perform an IAT for the first time during an experimental session. One explanation for this apparent discrepancy is that we focussed on novel attitudes whereas previous studies looked at overlearned attitudes and associations. It could be that it is easier to fake novel attitudes because it is not necessary to counteract well-established existing attitudes toward those same attitude objects. However, before strong conclusions can be drawn about this issue, new studies are needed in which faking attitudes toward novel and familiar attitude objects are compared directly.

Our results also demonstrate that participants can intentionally influence not only the magnitude of the IAT effect but also the direction of the effect. In the consistent fake condition, the IAT revealed more positive attitudes toward the category that was said to be good than toward the category that was said to be bad. In the inconsistent fake condition, however, a significant reversed effect was found. It is interesting to note that the effect in the inconsistent faking condition was in absolute terms smaller than the IAT effect in the consistent fake condition. Hence, it seems more difficult to fake attitudes that are opposite to the provided information than to fake attitudes that are consistent with the provided information even when the participants known that the information is made up. One possible reason for this finding is that participants in the inconsistent faking condition find it more difficult to determine or remember what it is that they are supposed to do. But it might also point to the fact that the mere presentation of information is sufficient to create implicit attitudes which then influence IAT performance in the same (consistent faking condition) or opposite (inconsistent faking condition) manner as intentional attempts to fake attitudes on the IAT. The latter account is, however, highly speculative.

Although participants could fake novel attitudes on the IAT, it is not clear how they succeeded in doing so. Steffens (2004; also see Fiedler and Bluemke, 2005) pointed at two possible strategies. First, participants could intentionally slow down or speed up responding in one of the IAT tasks. However, an inspection of the individual mean reaction times suggests that our participants did not adopt such a strategy, at least not in an obvious manner. In the consistent fake condition, only 4 out of 67 participants had an average reaction time of more than 1300 ms in the incompatible task. In the inconsistent fake condition, only 2 out of 61 participants on average responded more slowly than 1300 ms in the compatible task (see Table 1 for the range of reaction times). This range of reaction times is similar to that observed in many other IAT studies that were conducted at our laboratory. A second strategy discussed by Steffens is that participants selectively retrieve prior experiences with the attitude object that are consistent with the instructions. For instance, when asked to fake a positive

attitude toward Black people, participants might think of situations in which they interacted with Blacks in a positive manner. There is indeed evidence that strategies such as selective retrieval of memories can influence IAT effects (see Blair, 2002, for a review). However, participants could not have used this strategy in the present experiment because they had no prior experience with the (fictitious) social groups.

A third possibility is that participants intentionally imagined that they liked one social group and disliked another social group. A related strategy would be to intentionally label one group as good and another group as bad in line with the way they were asked to respond to the groups. As such, participants would explicitly associate the groups with the positive or negative images or with the concepts "good" or "bad." Such imagery or labelling could have had an impact on IAT effects. One could even argue that imagery and labelling can lead to genuine attitudes. If this is true, then perhaps our results are not problematic for the IAT because the observed IAT effect could have reflected genuine novel attitudes that were formed as a result of imagery or labelling. Even if this argument is correct, the fact remains that participants can influence IAT effects at will, be it by intentionally manipulating their (implicit) attitudes. It would still be problematic to use the IAT to study how implicit attitudes form spontaneously, that is, when participants do not attempt to please the researcher by showing the expected effects. It is possible that future research will identify performance variables (e.g., SD of reaction times) that allow one to determine whether a particular IAT score is due to faking. But until such variables are identified, the use of implicit measures as a way to prevent demand effects in studies on attitude acquisition should be avoided.

Appendix A. Instructions

A.1. First screen

In a recent study, it was examined how people form an opinion about social groups. Participants were told that the study was about two so-called ethnic groups that are called Luupites and Niffites. Luupites and Niffites can be distinguished by their name. All the names of Luupites contain two consecutive vowels and end on "lup" (e.g., Meesolup, Naanolup, Omeelup, Wenaalup). All the names of Niffites have two consecutive consonants and end on "nif" (e.g. Cellanif, Eskannif, Lebbunif, Zallunif). As you probably suspect, Luupites and Niffites do not really exist. They were invented by the researchers. The researchers told half of their participants a (made-up) story in which Luupites were depicted as "bad guys" and Niffites as "good guys." Other participants were told that Luupites are "good guys" and Niffites "bad guys." Afterwards, participants were asked to indicate on a scale from 1 to 10 how they thought and felt about the groups. They also had to complete a reaction time task that could be used to determine how positive or negative they were with regard to the groups. The results showed that the thoughts and feelings corresponded to the content of the story.

However, we suspect that the participants were aware of the aim of the experiment and based their answers on what, according to them, the researcher would like to find. In other words, we suspect that the participants intentionally tried to behave in a manner that was expected from them. If this is correct, then their responses are not a reflection of their true thoughts and feelings, but are purely due to a demand effect (that is, the participants chooses to behave in the manner that the researcher asks).

If the results are due to demand effects, then we should find the same results when we explain to people that all information about the groups is made up and thus wrong, but at the same time ask participants to respond in line with the expectations of the researcher. It is this hypothesis that we want to test in the current study.

A.2. Second screen

Note that the other half of the participants received identical instructions except that Niffites were described as good and Luupites as bad. The instructions that were different in the consistent and inconsistent faking condition are between brackets.

Here are the instructions of the current study: Imagine that a researcher tells you that Luupites are bad guys (they supposedly are violent, wild, malicious) and that Niffites are good guys (they supposedly are peaceful, civilised, and benevolent). The researcher expects that because of this information, you will have negative thoughts and feelings with regard to Luupites and positive thoughts and feelings with regard to Niffites. Of course, the information that the researcher gives you is entirely made up, but still the researcher wants you to answer in such a manner that you confirm his expectations. [CONSISTENT FAKING CONDITION: Your task is to do what the researcher expects.] [INCONSISTENT FAKING CONDITION: Your task, however, is to do the OPPOSITE of what the researcher expects.] Also remember very well what the researcher wants you to do (namely to respond positively toward NIFFITES and negatively toward LUUPITES). You will need this information again at several moments. On these moments, we will test your memory for the information of the researcher. So do not forget the instructions and expectations of the researcher (Niffites = good; Luupites = bad) [CONSISTENT FAKING CONDITION: and try to "fake" that you have these feelings] [INCONSISTENT FAKING CONDITION: but try to "fake" that you have the opposite feelings]. As in the original research, you will have to give judgments by clicking on a number of a scale and you will be given a reaction time task. You will get more information about this in a minute. After you have pressed the key "v," the actual experiment will start. From that moment onwards, act as if your feelings toward

NIFFITES and LUUPITES are [CONSISTENT FAKING CONDITION: like the researcher wants them to be] [INCONSISTENT FAKING CONDITION: opposite to what the researcher wants].

A.3. Third screen

In the reaction time task, words will appear on the screen one by one. The meaning of the words is always important. Each word belongs to one of four possible categories:

(1) positive words (e.g., happy)
(2) negative words (e.g., false)
(3) names of LUUPITES
(4) names of NIFFITES

You need to press on the left (q) or right (m) key depending on the category to which the presented word belongs. Which categories are assigned to which key differs from phase to phase. The categories that are assigned to the left key will appear in the left top corner of the screen; the categories assigned to the right key will appear in the right top corner of the screen. Try to respond as quickly as possible without making too many errors. This task will last approximately 10 min.

References

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242–261.

De Houwer, J. (2006a). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37*, 176–187.

De Houwer, J. (2006b). What are implicit measures and why are we using them. In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction*. Thousand Oaks, CA: Sage Publishers.

De Houwer, J., & Moors, A. (in press). How to define and examine the implicitness of implicit measures. In B. Wittenbrink (Ed.), *Implicit measures of attitudes*. Guilford Press.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: a review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853–869.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). San Francisco, CA: Academic Press.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: their meaning and use. *Annual Review of Psychology, 54*, 297–327.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Test. *Basic and Applied Social Psychology, 27*, 307–316.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731.

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

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*, 1464–1480.

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

Gregg, A. P., Banaji, M. R., & Seibt, B. (2006). Easier made than undone: the asymmetric malleability of automatic preferences. *Journal of Personality and Social Psychology, 90*, 1–20.

Kim, D. Y. (2003). Voluntary controllability of the Implicit Association Test. *Social Psychology Quarterly, 66*, 83–96.

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation, 34*, 203–217.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*, 413–417.

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition, 20*, 89–104.

Steffens, M. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51*, 165–179.

Zajonc, R. B. (1980). Feeling and thinking. Preferences need no inferences. *American Psychologist, 35*, 151–175.