Figure-Ground Asymmetries in the Implicit Association Test (IAT)

Klaus Rothermund and Dirk Wentura

Universität Trier and Universität Münster

Abstract. Based on the assumption that binary classification tasks are often processed asymmetrically (figure-ground asymmetries), two experiments showed that association alone cannot account for effects observed in the Implicit Association Test (IAT). Experiment 1 (N = 16) replicated a standard version of the IAT effect using old vs. young names as target categories and good and bad words as attribute categories. However, reliable compatibility effects were also found for a modified version of the task in which neutral words vs. nonwords instead of good vs. bad words were used as attribute categories. In Experiment 2 (N = 8), a reversed IAT effect was observed after the figure-ground asymmetry in the target dimension had been inverted by a previous go/nogo detection task in which participants searched for exemplars of the category "young." The experiments support the hypothesis that figure-ground asymmetries produce compatibility effects in the IAT and suggest that IAT effects do not rely exclusively on evaluative associations between the target and attribute categories.
Key words: Implicit Association Test (IAT), measurement of implicit attitudes, age stereotypes

„Figur-Grund-Asymmetrien" im Implicit Association Test (IAT)

Zusammenfassung. Der Implizite Assoziations-Test (IAT) besteht aus binären Klassifikationsaufgaben. Bei binären Entscheidungsaufgaben wird aber häufig auf die salientere der beiden Kategorien fokussiert. In zwei Experimenten konnte gezeigt werden, daß IAT-Effekte auf solche „Figur-Grund-Asymmetrien" zurückgehen. In Experiment 1 (N=16) wurde eine Standardvariante des IAT repliziert, bei der typisch alte vs. typisch junge Namen als Zielkategorien und positive vs. negative Wörter als Attributkategorien eingesetzt wurden. Allerdings wurde ein Kompatibilitätseffekt auch für eine modifizierte Version dieser Aufgabe gefunden, in der neutrale Wörter vs. Nichtwörter als asymmetrische, aber valenzneutrale Attributkategorie anstelle der positiven vs. negativen Wörter benutzt wurden. Im zweiten Experiment (N=8) fand sich ein umgekehrter IAT-Effekt, nachdem die Figur-Grund-Asymmetrie für die Zielkategorien invertiert wurde. Hierzu wurde vor dem IAT eine Detektionsaufgabe durchgeführt, bei der nach typisch jungen Namen gesucht wurde. Die Experimente stützen die Hypothese, daß Figur-Grund-Asymmetrien Kompatibilitätseffekte im IAT produzieren. Assoziationen zwischen Ziel- und Attributkategorien allein reichen nicht aus, um IAT-Effekte zu erklären.
Schlüsselwörter: Implicit Association Test (IAT), Messung impliziter Einstellungen, Altersstereotype

Recent research in social psychology has sought to identify and analyze cognitive processes that lie at the heart of certain phenomena. For example, it has been argued that attitudes, prejudices, and stereotypes arise from associative structures that are activated by situational or personal cues. Such cues, it is thought, increase the accessibility of potentially relevant information and evaluations about exemplars of a social category (Bargh, Lombardi, & Higgins, 1988; Devine, 1989; Fazio, in press; Fazio, Sanbonmatsu, Powell, & Kardes, 1986). Research has shown that an activation of category-related information and evaluations is automatic, i.e., without intent and sometimes even without conscious recognition of the effects of the activation (e.g., Bargh, Chaiken, Govender, & Pratto, 1992; Bargh & Pietromonaco, 1982; De Houwer & Eelen, 1998; Fazio et al., 1986; Greenwald, Draine, & Abrams, 1996). Greenwald and Banaji (1995) have termed these automatic activation processes "implicit cognition".

Recently, Greenwald, McGhee, and Schwartz (1998) introduced a new implicit measure to analyze associative cognitive structures, the Implicit Association Test (IAT). The IAT consists in a combination of two binary classification tasks, i.e., four stimulus categories are assigned to two responses. Two of the four categories represent the target concepts. Each target category is assigned to one of the two responses (e.g., insects vs. flowers: for insects, press the right key; for flowers, press the left key). The other two categories form the attribute dimension (e.g., valence). The two categories of the attribute dimension are assigned to the same two responses as the target categories (e.g., for pleasant words, press the right key; for unpleasant words, press the left key). After practising the binary classifications separately for the target and attribute categories, a *combined task* is performed. In the combined task, stimuli from all four categories are presented in random sequence and participants attempt to respond correctly. For each participant, the combined task is presented in two different blocks that contain different versions of the combined task. Before the second block, one of the response assignments is switched – typically that of the target categories (e.g., for insects, now press the left key; for flowers, now press the right key). An IAT effect is computed as the difference of the mean response times (RT) between the two versions of the combined task. If, for example, insects are associated with an unpleasant evaluation (and/or flowers with a pleasant one), the mean RT for the compatible block (i.e., insects and unpleasant words are assigned the same response) is subtracted from the mean RT for the incompatible block (i.e., insects and unpleasant words are assigned different responses). Typically, a difference score that is significantly above zero is found, i.e, participants are relatively faster in the compatible block. This result is interpreted as evidence for an association between the categories "insect" and "unpleasant" and/or "flower" and "pleasant".

Compared to other methods that are used in social cognition research to investigate implicit cognitive associations – e.g., semantic or evaluative priming – the IAT has some distinct advantages. Most importantly, effect sizes are much larger in the IAT compared to other response time paradigms (Greenwald et al., 1998, p. 1477). In addition, any target or attribute concept whatsoever can be used in the IAT, which makes it a very flexible research instrument. The availability of appropriate research tools (user-friendly programs) and the ease with which experiments can be conducted (relatively few participants and trials are needed) have led to a real boom in research with the IAT (as is documented, e.g., in this Special Issue).

Despite these advantages, the processes and mechanisms mediating between associations and an IAT effect are not yet fully understood. Why might one suppose that compatibility effects in the IAT reflect automatic cognitive associations between target and attribute categories? Greenwald et al. (1998) offered the following thought experiment: Imagine a task similar to the IAT, in which female and male faces as well as female and male names are assigned to the same responses ("hello" vs. "goodbye"). In this case, it is obvious that faster classification responses will emerge for conditions with a consistent response assignment for faces and names (e.g., for female faces and female names, say "hello"; for male faces and male names, say "goodbye") compared to conditions with an inconsistent response assignment. They explained this phenomenon as follows:

"The expected difficulty of the experiment with the reversed discrimination follows from the existence of strong associations of male names to male faces and female names to female faces. ... The (assumed) performance difference between the two versions of the combined task indeed measures the strength of gender-based associations between the face and name domains." (Greenwald et al., 1998, p. 1464)

But this argument is not fully convincing. On the one hand, the assumed association between faces and names of the same sex is unclear and questionable. Why should the faces of strangers be associated with certain names? Of course, everyone knows that female names belong to female faces, but is this abstract piece of knowledge truly an association? An even more critical question is whether this kind of knowledge is responsible for the expected difference in response latencies for consistent and inconsistent response assignments. Such an effect can readily be explained by the fact that in the compatible condition, the task can be reduced to a simple binary classification (if the stimulus is female, say "hello"; if it is male, say "goodbye"). This, of course, is not possible in the inconsistent response condition.

Similar problems arise with respect to the interpretation of IAT effects for attitude- or stereotype-related target categories reported by Greenwald et al. (1998). To get a full understanding of the problems involved at least three questions must be discussed: (1) What exactly does the alleged association between target categories and evaluative attributes refer to? Do these associations exist for the exemplars of the target categories or for the abstract target concepts? (2) Which processes mediate between category associations and IAT effects? (3) How can one be sure that IAT effects are not caused by other processes that are unrelated to associations between target and attribute categories?

Ad (1). What is associated? Experiments by De Houwer (in press) and Neumann (1999) show that IAT effects are for the most part independent of the exemplar stimuli used. Neumann (1999) found an IAT effect when photos of black and white persons were classified as "German" vs. "foreign" (by German participants), but no such effect when the same photos had to be classified as "white" vs. "black". A comparable result was reported by De Houwer (in press). Half of the stimuli of the target categories "British" and "foreign" consisted of positive exemplars (e.g., Princess Diana, Albert Einstein), the other half consisted of negative exemplars (e.g., M. Thatcher, Pinochet). The experiment was conducted with British participants. In this case, the observed IAT effect (faster responses if "foreign" and "negative" were assigned the same response) was independent of the valence of the target stimuli. Apparently, IAT effects are not located at the level of individual exemplars but rather at the level of the global target concepts.

Ad (2). Which processes mediate between category associations and IAT effects? Conceding for a moment that IAT effects reflect associations between target and attribute categories: Which processes mediate between category associations and IAT effects? Careful consideration reveals that giving an answer to this question is not an easy task. In a recent paper, De Houwer (in press; see also De Houwer, Hermans, & Eelen, 1998, pp. 89–90) has proposed an account of association-based IAT effects. According to De Houwer, association-based IAT effects can be explained in terms of stimulus-response compatibilites and incompatibilities. These effects emerge on the basis of acquired meanings of the response keys: Previously neutral responses in the IAT task (e.g., press the left key) acquire the characteristics of the categories that are assigned to them, e.g., a response acquires a positive valence when it is assigned to a positive category. In compatible blocks of the IAT, the target and attribute categories assigned to each response are similar with regard to a certain feature (e.g., valence). Therefore, each response unequivocally acquires the common feature of the assigned categories. In incompatible blocks, however, each response acquires conflicting features (e.g., positive *and* negative valence) because the target and attribute categories assigned to each response are opposed with regard to this feature. Stimulus-response compatibility effects arise because stimuli that belong to a certain category automatically elicit a corresponding response tendency (e.g., stimuli that belong to a valent category automatically elicit responses with a corresponding valence, see also Wentura, Rothermund, & Bak, 2000). Therefore, in a compatible block, the correct response is automatically activated

in each trial. In incompatible blocks, on the other hand, "stimuli will (a) automatically activate the representation of the incorrect response and/or (b) automatically activate the representation of the correct response to a lesser extent than with compatible response assignments" (De Houwer, in press).

The account given by De Houwer (in press) is an elegant explanation of IAT effects in terms of stimulus-response compatibilites and incompatibilities. But does this account really refer to *associations* between target and attribute categories? From our point of view, associations between categories must not be equated with similarity or common features (think of the distinction between semantic and associative priming, e.g., Williams, 1996). For instance, low familiarity can be a common feature of two categories, although the categories are in no way associated. Alternatively, two categories might be associated without being similar with respect to any feature whatsoever (e.g., the categories might be typically enumerated together, think of "height and weight", "gender, race, and religious belief"). Therefore, although De Houwer's account can explain why it is easier to respond if the target and attribute categories that are assigned to the same response share a common feature (whatever that feature might be), it does not provide an account in terms of associations between target and attribute categories (unless they share a common feature).

Ad (3). What else might produce compatibility effects in the IAT? The crucial question, thus remains: Can compatibility effects in the IAT be unambiguously and exclusively attributed to automatic associations between target and attribute categories? Interpreting compatibility effects in the IAT as evidence for cognitive associations requires that these associations are not only sufficient but also *necessary* determinants of these effects. Even if it can be made plausible that target-attribute associations do produce corresponding IAT effects, this by no means justifies conclusions in the other direction. An observed IAT effect might still be due to some other characteristics that are partially or completely independent of target-attribute associations.

For example, Dasgupta, McGhee, Greenwald, and Banaji (2000) discussed whether IAT effects might be due to familiarity differences of the target category exemplars. They found an automatic preference for White Americans compared to African Americans even when familiarity of stimuli was controlled for. To the contrary, Brendl, Markman, and Messner (in press) argue that differences in the familiarity of target categories have a strong influence on IAT effects. They replaced the target category "flowers" in the well-known insects vs. flowers IAT with non-

words (Brendl et al., in press). In the original study by Greenwald et al. (1998, Exp. 1), response latencies were markedly faster when insects and unpleasant words were assigned to the same response. This effect was completely reversed in the experiment by Brendl et al. (in press). Given the results of Brendl et al. (in press), to defend the hypothesis that IAT effects are due to target-attribute associations, one would have to accept the implausible assumption that the category of non-words is even more strongly associated with unpleasant evaluations than the category of insects[1].

The apparent discrepancy between the studies of Dasgupta et al. (2000) and Brendl et al. (in press) can be attributed to the fact that whereas Dasgupta et al. (2000) were concerned with the characteristics of individual stimuli, i.e., the category *exemplars*, Brendl et al. (in press) were concerned with features of the *categories* (see above). This brings us to our main objective, a theoretical analysis of IAT effects in terms of category asymmetries.

An Alternative Account: Explaining IAT Effects by Figure–Ground Asymmetries

In the present paper, we will provide a general alternative account of IAT effects that is not based on associations between target and attribute categories. We argue that IAT effects can be interpreted in terms of figure-ground asymmetries. What is meant by this? Typically, it is assumed that a participant's behavior in a binary classification task is symmetrical with regard to the categories A and B: If the stimulus belongs to category A, the "category A" response key will be pressed; if the stimulus belongs to category B, the "category B" response key will be pressed. However, participants may solve the task by focusing on only one of the two categories, thus producing an attentional asymmetry. That is, a binary classification task can be compared to a visual search task. In a visual search task, participants have to respond with "yes" whenever the stimulus display contains a stimulus of the target category and with "no" whenever the stimulus display contains only stimuli of the distractor category (e.g., Wolfe, 1998). In this

regard, a binary classification task can be compared to a visual search task with a display set size of one stimulus. Instead of keeping both stimulus-response assignments equally accessible in working memory, the classification task is reduced to a unipolar search task which focuses on the elements of the "figure" category, i.e., the category that was implicitly chosen as the target category. If an exemplar of this target category is detected (i.e., a match occurred) the assigned response is executed, otherwise (i.e., no match occurred) the other response is emitted.

When will participants tend to behave in this way? An asymmetry is especially likely if the figure and ground categories differ in salience. For example, the exemplars of the figure category might "pop out" relatively to the exemplars of the ground category with regard to some visual feature (e.g., brightness, colour, or form). Alternatively, figure-ground asymmetries might originate from asymmetries in the category labels. For example, a common distinction in psycholinguistics refers to marked and unmarked language codes (e.g., Greenberg, 1966). Unmarked codes represent the common use of language (e.g., when comparing two objects with regard to their length, it is common to express the result as "X is *longer* than Y" rather than "Y is *shorter* than X"; i.e., "long" is the unmarked code whereas "short" is the marked code). When contrasting two categories, differences in the marking or conciseness of the category labels will often lead to a figure-ground asymmetry. In a classification task, the marked category will typically constitute the "figure", because it is uncommon and automatically arouses more attention. Similarly, figure-ground asymmetries might also result from a differential familiarity of the target categories. The less familiar category will attract more attention and will thus be used as the search figure. The hypothesis of attentional asymmetries between marked and unmarked, or familiar and unfamiliar stimuli is supported by research with the visual search paradigm. For instance, Wang, Cavanagh, and Green (1994) found that unfamiliar targets pop out among familiar distractors but not vice versa. In a similar vein, Johnston and Hawley (1994) report attentional effects of a "novel pop-out" and a "familiar sink-in" for new and recently presented stimuli, respectively.

In what way can figure-ground asymmetries explain compatibility effects in the IAT? If there is a figure-ground asymmetry both within the target categories and within the attribute categories, then – when compatible responses are required (with regard to the figure-ground asymmetry) – it is easy to simplify the simultaneous classification task of targets and attributes: Participants need only classify the stimulus according to its "figure" quality (i.e., sa-

---

[1] Nonsense poetry provides a good example of the implausibility of such an assumption. For instance, the poem "Das große Lalula" by Christian Morgenstern (from the collection "Galgenlieder" [Gallows Songs]) consists entirely of non-words. Nevertheless, reading the poem is in no way associated with negative feelings. If anything, reading it is associated with a feeling of genuine delight and enjoyment.

lience), regardless of whether it is a target or an attribute exemplar. In case of a figure exemplar, they execute the response that is assigned to the figure categories, otherwise they execute the other response. In this case, the simultaneous classification task is reduced to a simple binary decision task, e.g., for figures, press the left key, for non-figures, press the right key. This simplified mode of responding is not possible if the categories that represent the target figure and the attribute figure are assigned to different responses. To the contrary: For each stimulus, one must determine whether it belongs to the target or attribute dimension in order to identify the correct response (e.g., if the figure belongs to the target dimension, then press the left key).

This account of IAT effects is supported by recent findings by Mierke and Klauer (2001). In the combined task, the IAT resembles a typical task-switching design: Participants have to switch permanently between two binary classification tasks. A common finding with this type of design is that response latencies are much longer after a task switch than after a task repetition (e.g., Allport, Styles, & Hsieh, 1994; Meiran, 1996; Rogers & Monsell, 1995). Interestingly, Mierke and Klauer (2001) found a strong effect of task-switching in the incompatible version of the combined task but only a weak effect of task switching in the compatible version of the combined task. This pattern of effects is exactly what one would predict on the basis of the figure-ground asymmetry hypothesis: In the compatible version, the two binary classification tasks can be reduced to only one dimension, i.e., searching for figures (vs. non-figures). Therefore, it is no longer necessary to switch between target and attribute classification tasks and effects of task-switching should be reduced. In the incompatible version, on the other hand, a reduction of the two tasks onto one binary classification is impossible. Participants will always have to switch between the attribute and target classification tasks which should lead to significant switch-costs. But although the pattern of effects observed by Mierke and Klauer is in line with the figure-ground asymmetry hypothesis, it can also be explained by a more restrained model which assumes that a reduction in the complexity of the combined classification task requires that all stimuli can be classified on the basis of their valence (Mierke & Klauer, 2001).

In addition, most of the previously reported IAT effects can be explained by the figure-ground hypothesis. Standard IAT effects using valent target and attribute categories can be explained with reference to the fact that exemplars of negative categories automatically attract attention (Fox, Lester, Russo, Bowles, Pichler, & Dutton, 2000; Pratto & John,

1991). The negative categories will thus constitute the figures of the respective dimensions whereas positively valent categories will serve as the ground. In case of a response assignment that maps categories of the same valence to the same response, a simple search for figure stimuli can be used to solve the combined classification task. This is not possible with an incompatible response assignment.

Familiarity effects – if there are any, see above – can be explained with a similar logic. In this case, unfamiliar stimuli and categories will represent the figures of the search. Thus, the result obtained by Brendl et al. (in press) can be explained by the figure-ground hypothesis if one assumes that non-words do not match pre-existing cognitive patterns and therefore produce attentional orienting responses. When combined with "insects" as a second target category, "nonwords" will thus form the figure category which explains the reversal of the IAT effect in the Brendl et al. experiment. In this regard, it is not a counter argument that IAT effects observed by Dasgupta et al. (2000) were not markedly affected by the familiarity of the exemplar stimuli because figure-ground asymmetries might be induced primarily by category labels. This is evident in the study of Neumann (1999): Contrasting the category labels "white person" and "black person" does not produce a clear or homogeneous figure-ground asymmetry for German participants, because both labels are comparable with regard to familiarity or markedness for them. On the other hand, contrasting the category "German" with the category "foreign" yields a clear asymmetry with "foreign" as the figure and thus will produce IAT effects when positively and negatively valent words are used as attribute categories. The results reported by De Houwer (in press, see above) can be explained in the same way.

It is important to note that our theoretical account allows for IAT effects that might indeed reflect automatic preference for, e.g., white Americans, young people, etc. However, our argument is that even in this case IAT effects will not be due to *associations* between target and attribute categories (nor will they be due to associations between exemplars of these categories). Instead, such an effect will reflect independent category asymmetries within the target and attribute dimensions that allow for a simplification of the compatible combined task by focusing on the salient categories (figures). Another important point we would like to make is that valence or automatic preference is not the only source that leads to figure-ground asymmetries; hence, the conclusion that IAT effects must reflect automatic preference is not warranted.

As is evident by the arguments above, different theoretical accounts can be used to explain IAT ef-

fects (e.g., associations between target and attribute categories vs. figure-ground asymmetries). An unambiguous decision about the processes that produce IAT effects requires that the different explanatory factors are experimentally separated. In the following experiment, we therefore try to produce IAT effects using figure-ground asymmetries and rule out valent associations between target and attribute categories.

Experiment 1

In this experiment, we compare a standard version of the IAT with the target categories "old" and "young" and the attribute dimensions "good" and "bad" with a modified version of this IAT using the "word/nonword" dichotomy instead of "good/bad". The standard version is one of the most reliable IAT effects, that is, RTs are reliably faster if "old" and "bad" are assigned the same response (Nosek, Banaji, & Greenwald, 2000). An association account would predict compatibility effects only for the standard version but not for the modified version, because there is no association between the target category "old" and the attribute dimension "nonword" (or between "young" and "words"). However, according to the figure-ground asymmetry hypothesis, compatibility effects are predicted for both the standard version and for the modified version, because it is plausible to assume a figure-ground asymmetry for the "word/nonword" dichotomy as well (see, e.g., Wentura, 2000). Nonwords should become the figure in this dichotomy for most of the participants. We thus predict faster response latencies for response assignments in which the categories "old" and "nonword" are mapped onto the same response.

Another prediction relates to the interaction of task-switching effects with compatibility in the standard and modified versions of the IAT. According to the figure-ground asymmetry hypothesis, for both types of the IAT task, task-switching effects should be stronger in the incompatible condition because the task can be reduced to a one-dimensional binary classification task – detecting figures – in the compatible condition. To the contrary, according to the model proposed by Mierke and Klauer (2001), a reduction in the complexity of the combined task in the compatible blocks is only possible if target and attribute stimuli can be classified according to their valence which is not possible if neutral words and non-words are used as attribute stimuli.

Method

Participants. 16 students of psychology (13 female, 3 male; median age = 20 years) from the University of Trier, Germany, participated in the experiment.

Materials. For each category (target stimuli: young and old names; attribute stimuli in the standard version: good and bad words; attribute stimuli in the modified version: words and nonwords), 10 stimuli were selected (see Appendix for a complete list of the stimuli). Word stimuli of the modified IAT had neutral or no valence, and nonwords were created out of neutral words by changing two or three letters. The mean number of characters was comparable for stimuli of all categories (mean length varied between 5.4 [for old names] and 5.9 [for bad words]).

Design. Each participant completed the standard version of the old/young-IAT and the modified version of the task. Order of presentation of the IATs was counterbalanced across participants. Within each IAT and order condition, all possible (initial) assignments of categories to responses for the target and attribute categories were realized equally often. Half of these response assignments yielded compatible response assignments (same response for categories "old" and "bad", and "old" and "nonword", respectively), the other half yielded incompatible response assignments (same response for "old" and "good", and "old" and "word"). For each version of the IAT, half of the participants received compatible response assignments first and the other half of the participants received incompatible response assignments first. For half of the participants in each version of the IAT, response assignments were switched for the target categories whereas response assignments were switched for the attribute categories for the other participants. All possible factorial combinations were realized equally often. Response assignments were manipulated independently for both IATs. Additionally, approximately half of all trials in the compatible and incompatible blocks of each type of IAT represented task repetition trials, i.e., the same classification task (target or attribute classification) had to be performed as in the previous trial. The other half of the trials consisted of task switching trials, in which a classification task that was not performed in the previous trial had to performed (attribute classification after target classification or vice versa).

Procedure. Within each IAT, participants first received two practice blocks with only one classification dimension. In a first block, only the stimuli of the target categories had to be classified, whereas in a second block, the stimuli of the attribute categories of the respective IAT had to be classified. During the practice blocks, each of the target stimuli appeared

once. The order of stimulus presentation was randomized for each participant. In the third block, participants had to classify stimuli of the target and attribute categories simultaneously. During this block, each participant received each stimulus twice, yielding a total of 80 trials that were presented in an individually randomized sequence. Due to the randomization of the stimulus presentation, approximately half of the trials represented task repetition trials, whereas the other half of the trials represented task shifting trials. The first 20 trials of this sequence were presented as practice trials. In a fourth and fifth block, both practice blocks were repeated, inverting the response assignments of either the target categories or the attribute categories (see *Design*). In a sixth block, participants again received 80 trials of the simultaneous classification task with the first 20 trials serving as practice trials. After a short break, the second IAT was presented in exactly the same fashion as the first.

All stimuli were presented in white uppercase letters in the middle of a black computer screen. Category labels were constantly shown at the top right and top left corners of the display, indicating the assignment of categories to responses. Two keys on the computer keyboard ('D' → left, 'L' → right) were marked as the response keys. In each trial, stimuli remained on the screen until a response was registered. If an incorrect response was made, the stimulus remained on the screen and an error message appeared in red. The stimulus and error message disappeared after the correct response key had been pressed. The next stimulus appeared after a delay of 150 ms.

Results

Mean response latencies were calculated for compatible and incompatible response assignments separately for the standard and modified versions of the IAT. Erroneous responses (9.1%) and reaction times that can be considered outlier values (i.e., those values which are 1.5 interquartile ranges above the third or below the first quartile, Tukey, 1977; 7.3%) were excluded from further analysis. Since order of presentation did not interact with compatibility effects in both the standard and the modified version of the IAT (both $F < 2.58$, ns), we will report analyses for measures aggregated over this factor. A significance level of $\alpha = .05$ (two-sided) was chosen for all analyses throughout the text. Unless stated otherwise, reported effects were significant at that level.

Compatibility effects for the standard and modified version of the IAT. Main effects of compatibility were significant for the standard version, $t(15) =$

8.55, as well as for the modified version of the IAT, $t(15) = 4.24$ (see Figure 1). In the standard version of the IAT, response latencies were 114 ms (*SD* 54 ms, $d = 2.11$) faster for the compatible response assignments (same response for categories "old" and "bad"), in the modified version of the IAT, response latencies were 72 ms (*SD* 68 ms, $d = 1.06$) faster for the compatible response assignments (same response for categories "old" and "nonword"). The effect of compatibility was stronger for the standard version than for the modified version, $t(15) = 2.67$.
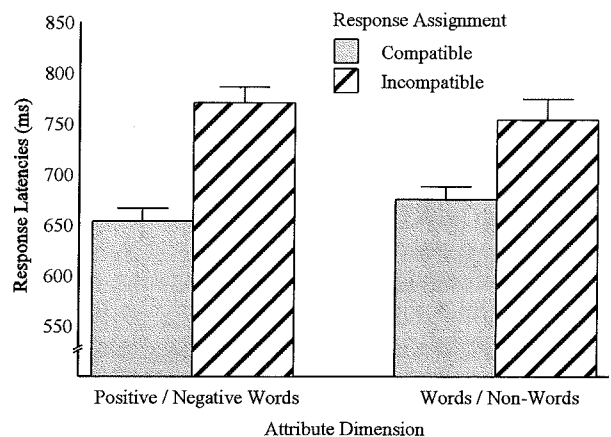

Figure 1. Mean response latencies and standard errors for the compatible and incompatible blocks of the standard and modified versions of the IAT (Experiment 1).

Effects of task shifting for the compatible and incompatible blocks of the standard and modified versions of the IAT. In an additional analysis, task shifting (task repetition trials vs. task switch trials) was introduced as an additional factor to analyze whether compatible and incompatible blocks differ with regard to task complexity. In this analysis, a significant main effect was observed for task shifting, $F(1,15) = 87.05$. Overall, response latencies were 74 ms faster for task repetition trials than for trials in which participants had to switch between the attribute and target classification trials (*SD* = 32 ms, $d = 2.33$). This main effect of task shifting was qualified by an interaction with compatibility, $F(1,15) = 15.89$ (see Figure 2). A task shifting effect of 98 ms was found in the incompatible blocks, $F(1,15) = 71.57$, $d = 2.11$, whereas a task shifting effect of 50 ms was found within the compatible blocks, $F(1,15) = 38.37$, $d = 1.55$ (see Figure 2). The three-way interaction of compatibility × task shifting × version of IAT was not significant, $F < 1$, indicating that the two-way interaction of task shifting × compatibility did not differ between the standard and the modified versions of the IAT (standard version: $F[1,15] = 13.52$; modified version: $F[1,15] = 6.79$).
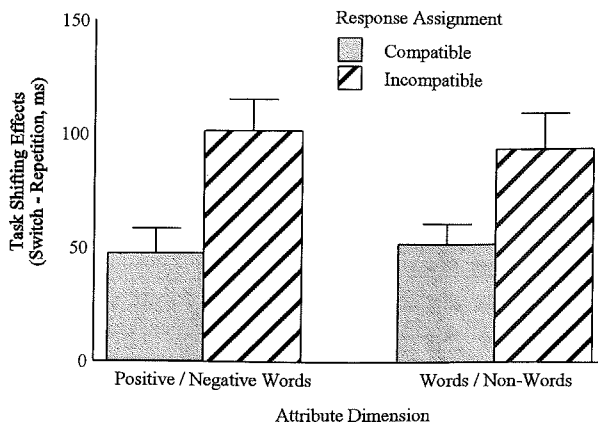
Figure 2. Mean effects of task shifting (and standard errors) for the compatible and incompatible blocks of the standard and modified versions of the IAT (Experiment 1).

Discussion

As predicted by both the association hypothesis and the figure-ground asymmetry hypothesis, a highly significant compatibility effect was observed in the standard version of the IAT using good and bad words as attribute categories. As predicted by the figure-ground asymmetry hypothesis only, a highly significant compatibility effect was observed for the modified version of the IAT using neutral words and non-words as attribute categories. This finding cannot be explained by the association hypothesis, because neutral words and non-words have no valence and have no association whatsoever with old and young names. This finding is convincing evidence for the existence of processes unrelated to associations between target and attribute categories which nevertheless produce strong response compatibility effects in the IAT paradigm. By implication, IAT effects cannot be taken as unambiguous evidence for the existence of such associations.

In the present experiment, compatibility effects were stronger for the standard version of the IAT. To explain this finding, one might assume that in the standard version, compatibility effects go back to a joint contribution of association or valence effects and figure-ground asymmetries. But this assumption is speculative. Differences in the strength of the effects might just as well reflect a difference in the salience of figure-ground asymmetries or other attributes on which the valent words of the standard version and the words and non-words of the modified version might differ.

Additional evidence for the figure-ground asymmetry hypothesis comes from the comparison of task shifting effects in the compatible and incompatible blocks of the combined task. Task shifting effects were significantly reduced in the compatible blocks which is in line with the assumption that IAT effects emerge as a result of a reduced task complexity in the compatible blocks (see Mierke & Klauer, 2001). As predicted by the figure-ground asymmetry hypothesis, this reduction in shift costs was found for both the standard and the modified version of the IAT. This finding rules out an explanation of the interaction effect that is restricted to valence as the mediating feature.

One might argue that although the first experiment provides sufficient evidence against the association hypothesis, it might not yet provide sufficient evidence in support of the figure-ground asymmetry hypothesis. Although the compatibility effect in the modified version of the task was predicted on the basis of the figure-ground asymmetry hypothesis, it might in principle be possible to explain the effect on the basis of other attributes relating to the word/nonword dimension or to the specific stimulus materials used that are independent of a figure-ground asymmetry (e.g., words and nonwords might be associated with response tendencies of approach or avoidance, respectively). Similar alternative interpretations might in principle be generated for the interaction effect of task shifting and compatibility in the modified version of the IAT.

Therefore, in the following experiment we tested the figure-ground asymmetry hypothesis more specifically by a direct experimental manipulation of figure-ground asymmetries that was independent of the categories and stimuli used. Alternative explanations of the compatibility effect in the modified IAT of the previous experiment relating to other attributes of the word/nonword dimension can be ruled out if the categories and stimuli are held constant. The following experiment should also produce additional evidence against the association hypothesis, since manipulating the direction of IAT effects independently of the categories and stimuli used is prima facie incompatible with an explanation of effects in terms of an association between categories or category exemplars.

Experiment 2

In the second experiment, only the standard version of the IAT was given to participants (old vs. young names as targets and bad vs. good words as attributes). However, the practice blocks requiring binary classifications for the target and attribute categories separately were modified in order to invert the typi-

cal figure-ground asymmetry for the target categories. Instead of a binary decision, participants had to carry out a go/nogo detection task during the practice trials, in which they had to press a key only for the young names (practice block 1) and only for the bad words (practice block 2). This manipulation should establish the figures "young" and "bad" for the following combined classification task that was presented in regular fashion, i.e., participants had to respond to each stimulus with one of two responses. The figure-ground asymmetry hypothesis now predicts an *inverted* compatibility effect, i.e., response latencies should be faster if the same response is assigned to the categories "young" and "bad". The association hypothesis, on the other hand, predicts the standard IAT effect, i.e., response latencies should be faster if the same response is assigned to the categories "old" and "bad".

Method

Participants. 8 students of psychology (2 female, 6 male; median age = 25 years) from the Universities of Trier and Münster, Germany, participated in the experiment.

Materials and Design. Materials and design were the same as in the standard IAT of Experiment 1.

Procedure. The practice blocks introducing the target stimuli (practice block 1) and the attribute stimuli (practice block 2) were now presented as go/nogo detection tasks. In the first block, only the category label "young" was shown at the top of the screen. Participants had to press either the left or the right key (according to the assignment schedule; see *Design* of Experiment 1) if and only if a young name was presented. Old names were displayed for 1000 ms and then disappeared. In addition, to enhance the asymmetry, we instructed participants to maximize a point account. That is, participants could gain up to five points by fast go-responses towards young names or lose up to five points by slow go-responses. After each response to a young name, gains were displayed in green and losses were displayed in red. The actual number of points was always shown above the middle of the screen. Erroneous go-responses to old names were commented with an error message in red and a loss of five points. A similar go/nogo task was conducted in the second practice block with bad and good words as stimuli. Participants had to press a key whenever they detected a bad word. In each of the first two blocks, each stimulus was presented four times, yielding a total of 80 trials for each go/nogo task. Trials were presented in an individually randomized order. In the third block, the standard combined classification task

of the IAT was conducted. In order to maintain the induced figure-ground asymmetries throughout the combined classification tasks, participants were instructed to focus primarily on the categories "young" and "bad". In order to prevent a relapse into the default figure-ground asymmetry triggered by the label "old", negated labels were shown for the ground categories in a somewhat smaller font size (i.e., "not young" and "not bad"). As in the go/nogo tasks, participants could gain up to five points by giving fast classification responses and lose up to five points by giving slow classification responses for stimuli of the categories "young" and "bad". As in the previous experiment, each stimulus was presented twice in the third block yielding a total of 80 trials, the first twenty of which were used as practice trials. In the fourth and fifth blocks, the go/nogo tasks were repeated with a reversed response assignment for the target category "young". In the sixth block, the combined classification task was repeated with the reversed response assignment. Presentation parameters were identical to the first experiment.

Table 1. Mean Response Latencies (Standard Errors in Parantheses) for Compatible and Incompatible Stimulus-Response Assignments under Reversed Figure-Ground Asymmetries for the Target Categories (Experiment 2)

Response Assignment	Standard Version (Old/Young, Good/Bad)
compatible[a]	852 (68)
incompatible[a]	779 (57)
IAT effect	−73 (29)

[a] Compatible response assignments denote that categories "young" and "bad" were assigned to different response keys; incompatible response assignments denote that categories "young" and "bad" were assigned to the same response key.

Results

Mean response latencies were calculated for all trials of blocks with compatible and incompatible response assignments. Erroneous responses (6.1%) and reaction times that can be considered outlier values (see Experiment 1; 7.3%) were excluded from further analysis. Mean response latencies are shown in Table 1. The main effect of compatibility was significant, $t(7) = -2.55$. Response latencies were now 73 ms (*SD* 81 ms, $d = .90$) faster for the *in*compatible response assignments (same response for categories "young" and "bad").

Discussion

The result of the second experiment yields strong support for the figure-ground asymmetry hypothesis. Inverting the figure-ground asymmetry for the target categories produces an inverted compatibility effect in spite of the fact that target categories and stimuli were identical to the standard version of the IAT in the first experiment. This effect cannot be explained by an association account of IAT effects since associations of target categories and stimuli were not changed by the manipulation. In fact, on the basis of the association hypothesis, an IAT effect in the opposite direction would have been predicted.

General Discussion

The experiments presented in this paper demonstrate a strong influence of figure-ground asymmetries on response compatibility effects in the IAT that are independent of associations between target and attribute categories. To avoid possible misunderstandings, we explicitly point out the implications we want to draw from the present findings as well as the implications we do not want to draw. Firstly, no attempt was made to rule out the possibility that associations between target and attribute categories can produce IAT effects. Although we argued in the introduction that there is still no satisfactory explanation of how associations influence compatibility effects in the IAT, we did not address this question in our experiments. We therefore do not claim to have shown that associations do not produce compatibility effects in the IAT nor do we claim to have shown that the IAT would be unable to detect such associations if they existed. By implication, we do not claim to have shown that all IAT effects must be due to figure-ground asymmetries. Our argument is simply that figure-ground asymmetries between the categories of the target and attribute dimensions do produce compatibility effects in the IAT. The experiments provide positive evidence that figure-ground asymmetries produce strong compatibility effects in the IAT, even in situations where associations or other factors related to the categories or to the exemplar stimuli cannot account for the effects.

A straightforward implication of this finding is that since figure-ground asymmetries can produce compatibility effects in the IAT, IAT effects in turn cannot be interpreted as unambiguous evidence for the existence of associations between target and attribute categories. In a given case, response compatibility effects in the IAT might reflect such associations, figure-ground asymmetries, or any combination of these and possibly other factors as well. Of course, demonstrating that IAT effects are due to figure-ground asymmetries in the target and attribute domains does not rule out the possibility that implicit cognitive associations exist between the target and attribute categories. They might exist, although they are not responsible for the response compatibility effects in the IAT. Alternatively, they might exist and produce – to some extent – the effects observed in the IAT. Or they might exist and influence figure-ground asymmetries which in turn influence the IAT. But of course, they may well *not* exist, despite a highly significant IAT effect. Without further information, an unambiguous interpretation of IAT effects is simply impossible. It should be kept in mind that this criticism is not restricted to a use of the IAT as a technique to demonstrate associative structures that are fairly universal – at least for a specified group of persons. *A fortiori*, this criticism applies to a usage of the IAT as a diagnostic device for the measurement of individual differences in the strength of such associations. By its very name, the IAT claims to be a psychological *test*. This kind of usage was obviously intended and propagated by Greenwald et al. (1998; see also Greenwald, Banaji, Nosek, & Bhaskar, 2000), who subscribe to the goal of replacing self-report scales measuring attitudes, prejudices or stereotypes by implicit measures that are immune to strategic processes of self-presentation. Notwithstanding the importance of this objective, the results of the present experiments cast some doubt on the construct validity of the IAT as a diagnostic tool for measuring interindividual differences in the strength of implicit cognitive associations. It might be a worthwhile endeavour for further research to try to disentangle various effects that produce response compatibility effects in the IAT or to develop versions of the IAT that can be more stringently related to implicit cognitive associations. However, any effort in that direction will have to overcome serious problems because the dimensions of valence, familiarity, and markedness are typically confounded (French, 1981; Hamilton & Deese, 1971). It might therefore be difficult to estimate effects of valence independently of these other factors. And even if some variant of the IAT might be able to solve these problems, it is doubtful whether it will still yield superior estimates in comparison to other techniques designed to analyse implicit cognitive evaluations and associations which are prima facie insensitive to artifactual effects of figure-ground asymmetries (e.g., evaluative priming, Fazio et al., 1986, affective Simon effects, De Houwer & Eelen, 1998, or the emotional Stroop, Williams, Mathews, & MacLeod, 1996). In this regard, it is noteworthy that the perhaps most promising technique to measure implicit evaluations – a response-window variant of the evalua-

tive priming paradigm – was invented by Greenwald and colleagues as well (see Draine and Greenwald, 1998; Greenwald et al., 1996; see also Musch, 1999; Otten & Wentura, 1999).

A second implication of our findings is somewhat more speculative. Although we do not claim that figure-ground asymmetries are the only possible source of compatibility effects in the IAT, we nevertheless believe that an account of IAT effects in terms of figure-ground asymmetries is an important one that sheds some light on the interpretation of other findings with the IAT. As was already mentioned in the introduction, most of the previous IAT studies used target and attribute categories that display a figure-ground asymmetry. This assumption is plausible at least for those dimensions that contain a negative or an unfamiliar category pole, because stimuli of these categories should automatically capture attention (Fox et al., 2000; Johnston & Hawley, 1994; Pratto & John, 1991; Wang et al., 1994). Attentional asymmetries might be a widespread phenomenon when investigating social categories (e.g., ethnic or stereotyped groups, ingroup-outgroup dichotomies, self- vs. other-related pronouns). In all of these cases, compatibility effects in the IAT can be strongly influenced by figure-ground asymmetries. The figure-ground asymmetry hypothesis might also help to explain the effects of contextual factors (e.g., priming or persuasive messages) on IAT effects (e.g., Kühnen, Schießl, Bauer, Paulig, Pöhlmann, & Schmidhals, 2001; Richter, Plessner, & Wänke, 2000). Context information might sometimes switch the focus of attention which can explain a change in the direction of the IAT effect. For example, activating a stereotype of East Germans might induce an attentional focus on the category "East German". Setting the focus on "East German" as the figure-category could explain why IAT effects for this category were biased in a negative direction (see Kühnen et al., 2001). The figure-ground asymmetry hypothesis also can account for some more subtle findings relating to the nature of the mediating processes underlying IAT effects. A case in point are the differences in the effects of task shifting in the compatible and incompatible blocks of the IAT (Mierke & Klauer, 2001). According to the figure-ground asymmetry hypothesis, the reduction in task shifting effects in the compatible blocks can be attributed to a reduction in task complexity if participants categorize stimuli of both target and attribute categories simply as "figure" vs. "ground". As was shown in the first experiment, a reduction in task switching effects in the compatible blocks was not restricted to the standard version of the IAT but also emerged in a modified version of the task that did not use valent stimuli or categories on the attribute dimension.

In our opinion, the figure-ground asymmetry hypothesis can account for a wide range of findings with the IAT. Future research might use this explanation of compatibility effects in the IAT in terms of attentional asymmetries within the target and attribute categories as a reference point against which explanations in terms of valence or associations can be tested.

References

Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and nonconscious information processing* (pp. 421–452). Cambridge, MA: MIT Press.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology, 62*, 893–912.

Bargh, J. A., Lombardi, W. J., & Higgins, E. T. (1988). Automaticity of chronically accessible constructs in person × situation effects on person perception: It's just a matter of time. *Journal of Personality and Social Psychology, 55*, 599–605.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology, 43*, 437–449.

Brendl, M., Markman, A., & Messner, C. (in press). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*.

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

De Houwer, J. (in press). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*.

De Houwer, J., & Eelen, P. (1998). An affective variant of the Simon paradigm. *Cognition and Emotion, 12*, 45–61.

De Houwer, J., Hermans, D., & Eelen, P. (1998). Affective Simon effects using facial expressions as affective stimuli. *Zeitschrift für Experimentelle Psychologie, 45*, 88–98.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General, 127*, 286–303.

Fazio, R. H. (in press). On the automatic activation of associated evaluations. *Cognition and Emotion*.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition and Emotion, 14*, 61–92.

French, P. L. (1981). Constructing comparative sentences: Linguistic marking and affect. *Journal of Psycholinguistic Research, 10*, 529–536.

Greenberg, J. H. (1966). *Language universals*. The Hague: Mouton.

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

Greenwald, A. G., Banaji, M. R., Nosek, B., & Bhaskar, R. (2000). *Implicit Association Test*. [WWW document]. URL http://www.yale.edu/implicit.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science, 273*, 1699–1702.

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.

Hamilton, H. W., & Deese, J. (1971). Does linguistic marking have a psychological correlate? *Journal of Verbal Learning and Verbal Behavior, 10*, 707–714.

Johnston, W. A., & Hawley, K. J. (1994). Perceptual inhibition of expected inputs: The key that opens closed minds. *Psychonomic Bulletin and Review, 1*, 56–72.

Kühnen, U., Schießl, M., Bauer, N., Paulig, N., Pöhlmann, C., & Schmidhals, K. (2001). How robust is the IAT? Measuring and manipulating implicit attitudes of East- and West-Germans. *Zeitschrift für Experimentelle Psychologie, 48* (this issue), 135–144.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1423–1442.

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für Experimentelle Psychologie, 48* (this issue), 107–122.

Musch, J. (1999). *Affektives Priming: Kongruenzeffekte bei der evaluativen Bewertung* [Affective priming: Congruency effects in evaluation]. Unpublished dissertation, University of Bonn, Germany.

Neumann, R. (1999). *Erleichterung oder Inhibierung: Wie funktioniert der Implizite Assoziationstest (IAT)* [Facilitation or inhibition: How does the Implicit Association Test (IAT) work]? Poster presented at the 7th Meeting of the Division of Social Psychology, Kassel, Germany.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2000). *Harvesting implicit group attitudes and stereotypes from a demonstration website*. Unpublished Manuscript.

Otten, S., & Wentura, D. (1999). About the impact of automaticity in the Minimal Group Paradigm: Evidence from affective priming tasks. *European Journal of Social Psychology, 29*, 1049–1071.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380–391.

Richter, L., Plessner, H., & Wänke, M. (2000). *Einstellungen und Verhalten gegenüber No-Name- und Markenprodukten* [Attitudes and behaviour related to no-name and brand products]. Paper presented at the 42nd Meeting of the Division of Experimental Psychology, Braunschweig, Germany.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207–231.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception and Psychophysics, 56*, 495–500.

Wentura, D. (2000). Dissociative affective and associative priming effects in the lexical decision task: Yes vs. no responses to word targets reveal evaluative judgment tendencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 456–469.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach and avoidance-related social information. *Journal of Personality and Social Psychology, 78*, 1024–1037.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*, 3–24.

Williams, J. N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology, 8*, 113–161.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–73). Hove, UK: Psychology Press.

Klaus Rothermund

Department of Psychology
University of Trier
D-54286 Trier
E-Mail: rothermu@uni-trier.de

Dirk Wentura

University of Münster
Psychologisches Institut IV
Fliednerstr. 21
D-48149 Münster
E-Mail: wentura@psy.uni-muenster.de

Appendix

Complete List of Stimuli used in the Experiments

Old Names (Exp. 1 and 2)	Young Names (Exp. 1 and 2)	Positive Words (Exp. 1 and 2)	Negative Words (Exp. 1 and 2)	Neutral Words (Exp. 1)	Neutral Non-Words (Exp. 1)
WALTER	TOBIAS	PARADIES (paradise)	ABGAS (exhaust fumes)	TISCH (table)	MUKET
OTTO	MARIO	HUMOR (humor)	VERLUST (loss)	PAPIER (paper)	TIMPF
JOSEF	SASCHA	LIEBE (love)	ANGST (anxiety)	FINGER (finger)	BALORT
ERICH	JENS	SOMMER (summer)	GEWALT (violence)	GABEL (fork)	FARST
HEINZ	PATRICK	FRIEDEN (peace)	KRIEG (war)	PUMPE (pump)	NIKAM
GERDA	KIRSTEN	FREUND (friend)	ELEND (misery)	BEGRIFF (concept)	SEDLOR
HILDE	JULIA	URLAUB (vacation)	SCHMERZ (pain)	BECHER (cup)	SCHESOL
MAGDA	JENNY	SANFT (gentle)	EINSAM (lonely)	FLACH (flat)	AGDAT
HEDWIG	SONJA	TREU (faithful)	GRAUSAM (cruel)	KARIERT (chequered)	RESCHLET
WALTRAUD	JASMIN	GESUND (healthy)	GEIZIG (miserly)	ECKIG (angular)	LESURT