Journal of Personality and Social Psychology
2003, Vol. 85, No. 6, 1180–1192

0022-3514/03/$12.00 DOI: 10.1037/0022-3514.85.6.1180

Method-Specific Variance in the Implicit Association Test

Jan Mierke and Karl Christoph Klauer
University of Bonn

The Implicit Association Test (IAT; A. G. Greenwald, D. E. McGhee, & J. L. K. Schwartz, 1998) can be used to assess interindividual differences in the strength of associative links between representational structures such as attitude objects and evaluations. Four experiments are reported that explore the extent of method-specific variance in the IAT. The most important findings are that conventionally scored IAT effects contain reliable interindividual differences that are method specific but independent of the measures' content, and that IAT effects can be obtained in the absence of a preexisting association between the response categories. Several techniques to decrease the impact of method-specific variance are evaluated. The best results were obtained with the D measures recently proposed by A. G. Greenwald, B. A. Nosek, and M. R. Banaji (2003).

The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is a simple experimental task used to measure the relative strength of associations between category-attribute pairs. On the basis of the assumption that many psychological phenomena such as attitude and self-concept can be conceptualized in terms of associative links between representational structures (e.g., Greenwald et al., 2002), the IAT has been applied in a variety of research domains such as racial attitudes (Dasgupta, McGhee, Greenwald, & Banaji, 2000; Greenwald et al., 1998; Ottaway, Hayden, & Oakes, 2001), attitudes towards homosexuality (Banse, Seise, & Zerbes, 2001), shyness (Asendorpf, Banse, & Mücke, 2002), marketing research (Maison, Greenwald, & Bruin, 2001), and dietary preferences (Swanson, Rudman, & Greenwald, 2001).

The standard IAT procedure (Greenwald et al., 1998) consists of two independent classification tasks that are performed in alternating order. Participants are asked to distinguish between exemplars of two target categories such as flowers and insects (categorization task), and to differentiate between stimuli high versus low on an attribute dimension, for example, between positive and negative stimuli (attribute task). In the standard procedure, the four response categories are assigned to two response keys in two complementary mappings. In one mapping, positive stimuli and flowers are mapped on the same response key, and negative stimuli and insects on the other. In a second mapping, in contrast, positive stimuli and insects are mapped on the same response, and negative stimuli and flowers on the other. The first mapping typically leads to faster and more accurate responses compared with the second for almost all participants, and it is often called "compatible mapping," whereas the second mapping is referred to as incompatible. The complementary response mappings are typically realized in separate phases of an IAT, and the IAT score is defined as the performance difference between these phases. According to Banaji (2001), this score "shows both the direction (positive vs. negative) of implicit attitude and the magnitude of the attitude (larger numbers reflecting larger differences)" (p. 124). The research reported in this article is concerned with the latter assertion that the strength of associative relations is reflected in the numerical size of the IAT score. On the basis of an account of IAT effects in terms of task-switching costs, we show that substantially unrelated IATs nevertheless share common method-specific variance. This implies that the magnitude of individual IAT scores also reflects stable interindividual differences that are unrelated to the purpose of measurement.

Psychometric Properties

Several recently published studies report results on the psychometric properties of IAT measures. Internal consistencies in the region of .80 have been obtained in the domains of implicit attitudes and self-esteem (Banse et al., 2001; Bosson, Swann, & Pennebaker, 2000; Cunningham, Preacher, & Banaji, 2001; Greenwald & Farnham, 2000; Greenwald & Nosek, 2001), anxiety (Egloff & Schmukle, 2002), and shyness (Asendorpf et al., 2002), among others. Retest correlations are somewhat smaller than internal consistencies, ranging between .27 (Cunningham et al., 2001) and .69 (Bosson et al., 2000). Retest correlations in the region of .60 seem to be the typical finding (Banse et al., 2001; Bosson et al., 2000; Greenwald & Farnham, 2000). These results indicate that properly designed IAT measures can capture a sufficiently high amount of systematic variance.

Investigations using behavioral criteria have recently provided a number of promising results concerning the validity of IAT measures. Asendorpf et al. (2002), for example, demonstrated a correspondence between an IAT measure and spontaneous overt behavior in the domain of shyness, Egloff and Schmukle (2002) demonstrated a relation between anxiety measured by the IAT and behavioral indicators of anxiety, and McConnell and Leibold (2001) found that IAT-measured prejudice was significantly correlated with certain parameters of nonverbal prejudiced behavior.

Jan Mierke and Karl Christoph Klauer, Psychological Institute, University of Bonn, Bonn, Germany.

This research was supported by German Research Foundation Grant KL 614 29-1 to Karl Christoph Klauer. We thank Jan De Houwer, Klaus Rothermund, Miguel Brendl, and Katja Ehrenberg for constructive comments and critique on an earlier version of this article, and Wera Hatje for help in collecting the data.

Correspondence concerning this article should be addressed to Jan Mierke, Psychological Institute, University of Bonn, Römerstrasse 164, Bonn D-53117, Germany. E-mail: jan.mierke@uni-bonn.de

However, a greater number of studies investigated convergent validity in the form of correlations between IAT scores and explicit measures of the same or a correlated construct, and most of these studies report low correlations, frequently even below .30 (e.g., Bosson et al., 2000; Cunningham et al., 2001; Egloff & Schmukle, 2002; Greenwald et al., 1998; Karpinski & Hilton, 2001; Nosek, Banaji, & Greenwald, 2002; Rudman, Greenwald, Mellott, & Schwartz, 1999; Swanson et al., 2001). Correlations above .30 have been obtained in a smaller number of studies (e.g., Asendorpf et al., 2002; Banse et al., 2001; McConnell & Leibold, 2001).

Accounts for the IAT Effect

Only a few studies have addressed the mechanisms underlying the IAT effect itself. Theoretical analyses can, however, provide fruitful insights on limitations and possibilities of IAT-based research, and several authors have recently suggested relevant models. A brief overview of the models proposed by Brendl, Markmann, and Messner (2001), De Houwer (2001), and Rothermund and Wentura (2001) is given in the following sections.

According to Brendl et al. (2001), the IAT effect reflects the result of a random walk process in which evidence is accumulated on a joint response-related decision dimension. The time required before a fixed response criterion is reached depends on whether all incoming information pushes an imaginary counter in the same direction. Instances of the target categories should have a lower net accumulation rate in the incompatible than in the compatible IAT condition, as information on the category membership and evaluation of a stimulus disagree in the former, but not in the latter, condition. As a consequence, the response criterion is shifted in the incompatible IAT condition (Brendl et al., 2001).

According to De Houwer (2001), the IAT effect is based on stimulus–response compatibility. The basic assumption in this model is that response keys acquire the meaning of the stimulus category they are assigned to. Compatibility between the meaning of a response key and stimulus features then facilitates responses with this key. This mechanism can explain the IAT effect, as compatibility between stimulus and response is consistently given in a compatible IAT phase, but not in an incompatible phase.

According to Rothermund and Wentura (2001), the IAT measures differences in the salience of stimulus categories. Figure-ground asymmetries within the target and attribute dimensions are the central explanatory concept of this account. The authors hypothesized that participants simplify the compatible task by recoding both classification tasks as figure-ground discriminations. A wide range of present IAT findings can be explained by assuming that asymmetries in salience are paralleled by asymmetries in valence or familiarity, even though, in principle, salience is dissociable from these latter constructs.

Task-Switching in the IAT

In the compatible condition of a typical IAT, the structure of the task provides participants with an overlapping attribute. In the compatible conditions of a flower–insect attitude IAT, for example, the attribute positivity is shared by positive adjectives and flowers, which are mapped to one response key, whereas negativity is shared by negative adjectives and insects, which are mapped to the second response key. Evaluating a flower or insect stimulus should thus lead to the same overt response as categorizing the stimulus (Mierke & Klauer, 2001). Consequently, the task-switching account assumes that participants derive their responses from the attribute shared by the target categories in the compatible condition. Because the process of deriving responses is thereby simplified, performance should be faster in this condition. However, responses cannot be derived from an overlapping attribute in the incompatible IAT condition; instead, attribute-related information needs to be ignored for target-category stimuli and processed for exemplars of the attribute categories. The central assumption of the task-switching account (Mierke & Klauer, 2001) is that this involves executive control processes, namely identifying and switching to the appropriate task set.

Even though there is some amount of debate concerning the nature of task sets and the process of switching between task sets (e.g., Allport, Styles, & Hsieh, 1994; Meiran, Chorev, & Sapir, 2000; Sohn & Anderson, 2001; Sohn & Carlson, 2000; Wylie & Allport, 2000), it can be concluded that task-set switching involves changing a complex of cognitive settings required for performing a given task, including “which attribute of the stimulus to attend to, which response mode and value to get ready, what classification of the relevant stimulus attribute to perform, how to map those classes to response values, with what degree of caution to set one’s criterion for response, etc.” (Monsell, Yeung, & Azuma, 2000, p. 252), and that the process of switching between task sets is associated with a performance cost (e.g., Meiran, 1996; Monsell et al., 2000; Rogers & Monsell, 1995).

A number of predictions derived from these assumptions have been tested in two earlier experiments (Mierke & Klauer, 2001). In both experiments the IAT procedure originally proposed by Greenwald et al. (1998) was slightly altered. Instead of presenting attribute and target stimuli in a strictly alternating order, which requires task switching on every trial, the task to be performed was repeated on a subset of trials within each block. Task-switching costs were then estimated by comparing performance for trials on which the task was repeated with trials on which the task had to be switched. As predicted, task-switching costs affected performance in the incompatible IAT condition, and had a significantly less pronounced effect in the compatible condition.

The purpose of the present article is to study how variance related to task switching affects the measurement of interindividual differences. A straightforward and plausible answer to this question is that task-switching costs play a *mediational* role, that is, stronger associative relations between target category and attribute result in higher costs for switching between ignoring and processing the associated attribute feature.

Yet, the literature suggests that task-switching performance is in itself subject to stable interindividual differences. There is evidence, for example, that task switches are performed faster by younger than by older persons (Kramer, Hahn, & Gopher, 1999; Kray & Lindenberger, 2000), and that persons with high fluid intelligence are faster in switching tasks than persons with lower fluid intelligence (Kray & Lindenberger, 2000). Thus, there is good reason to believe that there are stable interindividual differences in task-switching performance that are independent of the particular construct to be measured. According to the account by task-switching costs, such content-independent interindividual differences in task-switching performance should give rise to a stable but content-independent variance component in IAT scores. We refer to this component as *reliable contamination*, or *method-specific variance*.

In contrast, a third possibility is that of unreliable, random contamination, which would only marginally affect the estimation of psychometric properties by introducing error variance. The crucial difference between reliable and unreliable contamination is that task-switching components in IAT scores are not assumed to be stable across trials and measurement occasions under unreliable contamination, whereas reliable contamination will reliably inflict the same bias in an individual's IAT score on every measurement occasion irrespective of the measured constructs. Thus, reliable contamination can cause conflation of internal consistencies, stability indices, and even result in significant correlations between IAT measures that do not overlap with respect to content.

On the basis of the account just sketched, the IAT procedure may not be restricted to measuring the strength of preexisting associative links between concept nodes (Greenwald et al., 2002). Given appropriate material, framing, and task specifications, any IAT measure may be sensitive to various other types of relation, like asymmetries in salience, or contingencies between stimulus features. According to the task-switching account, performance will be slowed, whenever two stimulus–response mappings that specify conflicting responses for a subset of stimuli have to be applied in succession, in comparison with a condition in which the stimulus–response mappings agree for the complete set of stimuli.

The experiments reported below build on these considerations by implementing an IAT variant that is based on contingencies between stimulus features rather than on preexisting associations between categories. Specifically, Experiment 1A investigates whether a simple contingency between stimulus features is sufficient to produce IAT-like effects and which psychometric properties the resulting measures have. In Experiment 1B, the reliability of a standard attitude IAT is assessed, and Experiments 2 and 3 investigate whether this IAT as well as a new extraversion IAT correlate with the IAT variant explored in Experiment 1A, although the latter IAT is not related to the former ones in terms of content.

Experiment 1A

The aim of Experiment 1A was to test whether implicit association effects can be obtained with an IAT measure that is not based on a preexisting association between target categories and attributes. Instead the IAT was based on an experimentally imposed relation between superficial stimulus features of geometrical objects. Therefore, correlations between repeated applications of this IAT measure cannot be explained by interindividual differences in an underlying relation. However, on the basis of a reliable contamination hypothesis, systematic interindividual differences are expected nevertheless. It is predicted that internal consistencies and retest correlations will be well above zero.

Method

Participants of Experiment 1A successively performed two identical IATs using geometrical objects as material. The target categories to be distinguished were red objects and blue objects. The attribute to be judged was the size of the stimuli. The stimuli representing the attribute belonged to neither of the target categories, that is, they were neither red nor blue, but colored in one of three alternative colors (just as positive and negative stimuli in a flower–insect IAT are neither flowers nor insects). Similar to a conventional attitude IAT, the target categories were discrete (as are the categories flower and insect), whereas the attribute "size" was a continuous variable like valence. Obviously, there is no preexisting association between size and color of the objects that could account for interindividual differences in the measure.

A link between the target categories and the attributes was created experimentally by imposing a contingency between size and color of the objects: All objects belonging to the target category "red" were small, whereas all objects belonging to the target category "blue" were large. It should be noted that this closely resembles the structure underlying the material in a conventional attitude IAT: Stimuli belonging to the target category "insect" are typically negative, whereas stimuli belonging to the target category "flower" are positive. In the present experiment, the mapping is referred to as compatible if all large objects are mapped to one response key and all small objects to the other, irrespective of color. Participants performed the contingency-based IAT task twice in direct succession. Compatibility order (compatible vs. incompatible mapping first) for the first and second IAT was counterbalanced between subjects. Additionally, the complete response assignment was reversed for one half of the participants.

Participants. Twenty-five volunteers (17 women and 8 men) participated in exchange for partial course credit or a monetary gratification of 5 Euro (approximately $5 at that time). Their mean age was 25.76 years (SD = 4.52). All participants were University of Bonn, Bonn, Germany, students with different majors.

Material. Simple geometrical objects that differed in color (red, blue, yellow, green, and pink), size (large vs. small), and form (rectangles, triangles, and circles) were used as stimuli. Size and color of the stimuli were clearly distinguishable: The large stimuli were approximately two times larger than small stimuli. The form of the object was task irrelevant and manipulated randomly to generate variance in the stimulus sets. To impose a relation between the unrelated properties of size and color, all red objects were small and all blue objects were large. Samples of these stimuli are depicted in Figure 1.

Procedure. All blocks consisted of the sequential presentation of 48 single geometrical objects. The objects were sampled randomly with the restriction that objects from each set appeared with equal frequency. Objects representing the attribute and target categories appeared in random order, resulting in trials that required a task switch and trials in which the task of the previous trial was repeated. The presentation of a new object was initiated 800 ms after the participants' response. Responding was allowed as soon as the stimulus was visible. Participants were instructed to categorize an object as either red or blue, whenever a stimulus belonged to one of these color categories, and to judge the size of all objects (small vs. large) not belonging to one of the target categories. Participants were told to respond to each stimulus as rapidly as possible while avoiding errors. They started the upcoming block by sequentially pressing the two response keys. After a short countdown, the block was initiated.

In total, 28 blocks were to be completed, 14 in each of two identical IATs. Performance data were recorded for every trial. The experiment started with two training phases consisting of two blocks each. In the first phase, the target discrimination task was practiced. The attribute judgment was trained in the second phase. The training blocks were followed by four combined blocks, in which both tasks were mixed and were mapped either compatibly or incompatibly, depending on the order-balancing condition. The remaining six blocks of the first IAT consisted of two training blocks and four combined blocks, for which the compatibility of response mapping was switched. At the beginning of each block, participants were informed about the object categories that were to appear in the upcoming block and their assignment to the response keys. The second IAT started

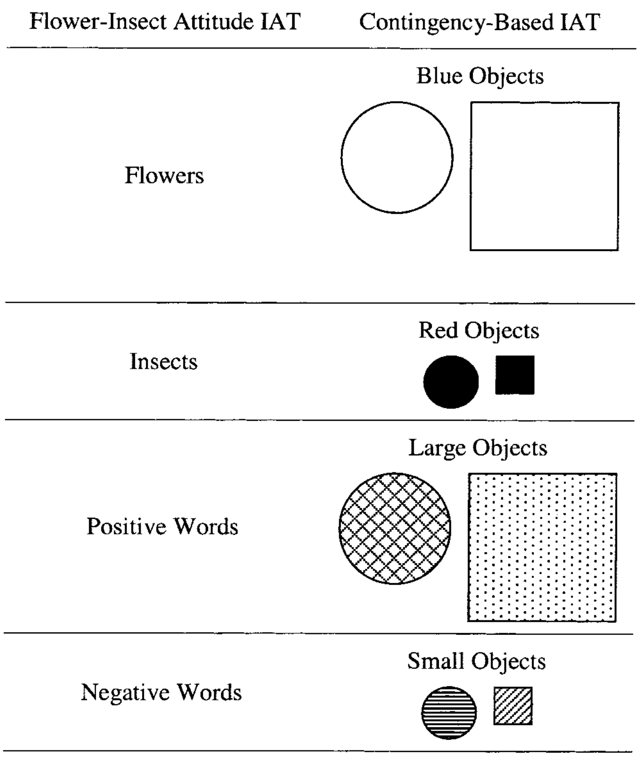

Figure 1. Sample material used in the contingency-based Implicit Association Test (IAT). Different fill-patterns represent different colors. Note that all blue objects are large, and all red objects are small.

directly after the first, and followed the same procedure. Instructions for the second IAT were shorter, however, to avoid redundancies.

Results

All trials with incorrect responses (6.3%) were discarded from the analysis. Trials with response latencies below 100 ms or above 3,000 ms (0.3%) were recoded to 100 ms and 3,000 ms, respectively. Furthermore, the first two trials of each block were removed. Response latencies were aggregated for each participant and in each cell of the within-participants design defined by the factors (a) compatibility of mapping, (b) first versus second IAT, and (c) task switch versus task repetition. IAT effects were computed for each factorial combination. A trial was defined as a task-repetition trial if the directly preceding trial required performing the same task as the current trial and as a task-switch trial otherwise.

Mean IAT scores were positive for the first (M = 374 ms, SD = 155) and the second (M = 281 ms, SD = 75) IAT measures, and ranged between 132 ms and 687 ms. To calculate internal consistencies, trials from the combined phases of each IAT were randomly assigned to one of four subtests. Cronbach's alphas were then calculated for the IAT scores obtained in these subtests. They were high for the first IAT (α = .95), and satisfactory for the second (α = .79). As expected, the correlation between the identical IAT measures was significant (r = .66, $p < .01$).

Analysis of variance (ANOVA). A 2 (IAT) × 2 (task switch vs. task repetition) × 2 (compatibility order) × 2 (mirrored stimulus–response mapping) mixed factorial ANOVA was performed on the IAT effects with repeated measures on the first two factors. To simplify the exposition, mapping compatibility was not entered as a separate factor. Instead, the analyses were based on IAT scores directly.

A main effect of task switching, $F(1, 21) = 40.41$, $p < .01$, revealed that the IAT effect was significantly larger for task-switch than for task-repetition trials. There was a smaller IAT effect in the second than in the first IAT, $F(1, 21) = 21.29$, $p < .01$. Unexpectedly, the order of the mapping conditions, $F(1, 21) = 4.57$, $p < .05$, mirrored response assignment, $F(1, 21) = 6.39$, $p < .05$, and an interaction of these factors, $F(1, 21) = 6.84$, $p < .05$, reached significance, showing that the IAT effect is larger if the compatible condition is to be performed first, and if small-sized objects have to be responded to with the dominant hand. As there were no hypotheses concerning these effects, they will not be discussed further. The mean IAT scores for task-switch and task-repetition trials can be found in Table 1.

Discussion

As predicted, the results of Experiment 1A show an implicit association effect in the absence of a preexisting association. The significant correlation between test scores on the two measurement occasions indicates that a common factor underlies the measurement. Interindividual differences in the strength of association between size and color of the stimuli may explain these effects, but it seems unlikely that such associations existed before the experiment. Finally, an ANOVA confirmed the basic prediction of the task-switching model, namely that IAT effects are significantly larger for trials requiring a task switch. This effect reflects the predicted asymmetry in task-switching costs (Mierke & Klauer, 2001; Rothermund & Wentura, 2001).

If it is assumed that stimulus sets or response categories differed with respect to their salience, the results of Experiment 1A can also be explained by the figure-ground asymmetry model proposed by Rothermund and Wentura (2001). Unlike most applied IATs, the contingency-based IAT allows reversing the direction of the rela-

Table 1
Mean IAT Effects in Experiments 1A, 1B, 2, and 3, as a Function of Task-Switching

	Task-switch trials		Task-repetition trials	
IAT measure	*M*	*SD*	*M*	*SD*
Geometry IAT (Exp. 1A, n = 24)	393	120	273	61
Geometry IAT (Exp. 2, n = 67)	452	178	292	128
Geometry IAT (Exp. 3, n = 81)	469	209	282	141
Flower–insect IAT (Exp. 1B, n = 25)	187	102	104	76
Flower–insect IAT (Exp. 2, n = 67)	267	182	120	89
Extraversion IAT (Exp. 3, n = 81)[a]	204	216	95	107

Note. IAT scores are based on mean untransformed response latencies for nonerror trials. Response latencies below 100 ms and above 3,000 ms were recoded to the respective values. IAT = Implicit Association Test; Exp. = Experiment.
[a] Extraversion IAT scores refer to the absolute, unsigned values in this table.

tion without affecting salience. We thus replicated Experiment 1A and reversed the contingency between size and color of the stimuli for one half of the participants, that is, all red objects were large and all blue objects were small for these participants. According to the figure-ground model, the IAT effects should not be affected by reversing the contingency, because switching the contingency between size and color should not affect preexisting salience asymmetries. A total of 24 persons participated in the replication study. Compatibility was coded in the same manner as in Experiment 1A, that is, the conditions in which the categories “red” and “small” share a response key were coded as compatible, irrespective of the underlying contingency. Whereas the mean IAT score was positive for the condition replicating the contingency of Experiment 1A ($M = 281.71$ ms, $SD = 218.18$), the effects were reversed ($M = -335.50$ ms, $SD = 169.64$) for the condition with switched contingency, $F(1, 16) = 52.93$, $p < .01$. This finding is incompatible with the figure-ground model (Rothermund & Wentura, 2001), because IAT effects were reversed even though the salience of response categories remained unchanged. This does not rule out, however, that figure-ground asymmetries are capable of causing IAT effects under other circumstances.

Experiment 1B

If the correlations and internal consistencies found in Experiment 1A are due to reliable contamination, a similar variance component should also affect more meaningful IATs. As a precursor for testing this prediction in Experiment 2, Experiment 1B was conducted. The purpose of Experiment 1B was to test whether the flower–insect attitude IAT used by Mierke and Klauer (2001) has a sufficiently high reliability to be used in the context of assessing interindividual differences.

Method

Participants of Experiment 1B performed a flower–insect attitude IAT twice in direct succession to assess its psychometric properties. The procedure and design of Experiment 1B were identical to Experiment 1A. Instead of a contingency-based IAT with geometrical objects, a conventional flower–insect attitude IAT was to be performed. More specifically, flower and insect names replaced the red and blue objects, whereas positive and negative stimuli were used instead of the objects instantiating the attribute categories large and small. All other aspects of the procedure remained unchanged.

Participants. A total of 24 persons (14 women and 10 men) participated in the experiment. Their mean age was 22.75 years ($SD = 2.74$). All participants were University of Bonn students with different majors and either received partial course credit or a monetary gratification of 5 Euro (approximately $5 at that time) for their participation.

Material. The material of Experiment 1B consisted of the 96 words referring to insects, flowers, positive objects, and negative objects, already used by Mierke and Klauer (2001). The words were matched in quadruples that were selected to be maximally similar on three criteria, namely the number of characters, an estimation of the word’s frequency of use based on the Celex lexical database (Celex, 1995), and a rating of the word’s valence. Details on this selection procedure can be found in Mierke and Klauer (2001).

Results

All trials with incorrect responses (6.7%) were discarded from the analysis. Trials with response latencies below 100 ms or above 3,000 ms (0.7%) were recoded to 100 ms and 3,000 ms, respectively. The first two trials of each block were also discarded.

Mean IAT scores were positive for the first ($M = 187$ ms, $SD = 116$) and the second ($M = 102$ ms, $SD = 70$) IAT measures, and ranged between -44 ms and 389 ms. The valid experimental trials of each IAT were again randomly assigned to four subtests. Cronbach’s alphas were calculated for the IAT scores obtained in these subtests. They were sufficiently high for the first and second IATs ($\alpha = .88$ and $\alpha = .82$, respectively). As predicted, the correlation between the measures was found to be significant ($r = .53$, $p < .01$).

ANOVA. Again, a 2 (IAT) × 2 (task switch vs. task repetition) × 2 (compatibility order) × 2 (mirrored stimulus–response mapping) mixed factorial ANOVA was performed on the data with repeated measures on the first two factors. The analysis revealed a main effect of task switching, $F(1, 20) = 38.25$, $p < .01$, that was due to larger IAT effects for task-switch than for task-repetition trials. The IAT effects were smaller when the IAT was performed for the second time, $F(1, 20) = 16.64$, $p < .01$. No other main effect or interaction gained significance. The mean IAT scores for task-switch and task-repetition trials can be found in Table 1.

Discussion

The pattern of results obtained in Experiment 1B closely resembles that of Experiment 1A. Stability and internal consistency of the flower–insect IAT were found to be nearly identical to those obtained with the geometric material. The flower–insect material seems to produce a satisfactory amount of systematic interindividual difference in the IAT scores. An ANOVA confirmed the predictions of the task-switching model, that is, IAT effects calculated on the basis of task-repetition trials were significantly smaller than those calculated on the basis of task-switch trials.

Experiment 2

A remarkable, nontrivial consequence of reliable method-specific variance is that it predicts correlations between IAT measures even when they do not overlap with respect to content. To directly test this prediction, participants of Experiment 2 successively completed two IAT measures, one with the flower–insect material, the other with the geometrical objects used in Experiment 1A. Because the contingency-based IAT measure is not related to the preference of flowers over insects, obtaining a correlation between these measures cannot be explained by interindividual differences in the measured construct. A positive and significant correlation would thus strongly indicate reliable contamination of the flower–insect IAT. If task-switching costs mediate valid variance, however, the measures ought to be uncorrelated.

Method

Each participant performed a contingency-based IAT with geometrical objects and a conventional flower–insect attitude IAT. Procedure and design of the IATs were identical to those of Experiments 1A and 1B. The order in which the IAT measures were administered was counterbalanced across participants as was the order of compatible and incompatible phases.

Participants. A total of 67 persons (47 women and 20 men) participated in the experiment. Their mean age was 25.67 years ($SD = 5.91$). All participants were University of Bonn students with different majors and either received partial course credit or a monetary gratification of 5 Euro (approximately $5 at that time) for their participation.

Materials. The materials of Experiment 2 consisted of the 96 words referring to insects, flowers, positive objects, and negative objects, used in Experiment 1B, and the geometrical objects used in Experiment 1A.

Results

All trials with incorrect responses (7.7%) were discarded from the analysis. The first two trials of each block were removed as well. Again, trials with response latencies below 100 ms or above 3,000 ms (0.9%) were recoded to 100 ms and 3,000 ms, respectively.

The mean IAT effect was 191 ms ($SD = 121.21$) for the attitude IAT and 366 ms ($SD = 143$) for the contingency-based IAT measure. Internal consistencies of the two IAT measures were estimated by computing Cronbach's alphas as before. The internal consistencies for IAT scores based on these subtests were sufficiently high for the attitude IAT ($\alpha = .88$), and high for the contingency-based IAT measure ($\alpha = .93$). As predicted, the correlation between the two substantially unrelated IAT measures was found to be significant ($r = .39$, $p < .01$).

Structural equation analysis. Additionally, the trials of the IAT measures on both measurement occasions were randomly split into two test halves. IAT scores were then calculated for each test half and entered into a structural equation analysis. The model depicted in Figure 2 was used for this analysis. This model assumes that two distinct, yet correlated latent variables (τ_a and τ_g) underlie the observed covariance matrix. The manifest variables IAT$_{a1}$ and IAT$_{a2}$ refer to the test halves of the flower–insect attitude IAT, and IAT$_{g1}$ and IAT$_{g2}$ refer to the test halves of the contingency-based IAT. Residuals for these variables are denoted by ε_{a1}, ε_{a2}, ε_{g1}, and ε_{g2}. All factor loadings were restricted to be equal within, but not across, the two IAT measures. The resulting equation model fit the data well, $\chi^2(5, N = 67) = 6.47$, $p = .26$. As expected, given high internal consistencies, a large portion of the variance within each IAT is true-score variance. The correlation between true scores for the two IAT measures was .42 with a standard error (*SE*) of .11. The correlation differed significantly from zero, $t(66) = 3.74$, $p < .01$. The covariance matrix is given in Table 2, and standardized parameter estimates can be found in Figure 3.

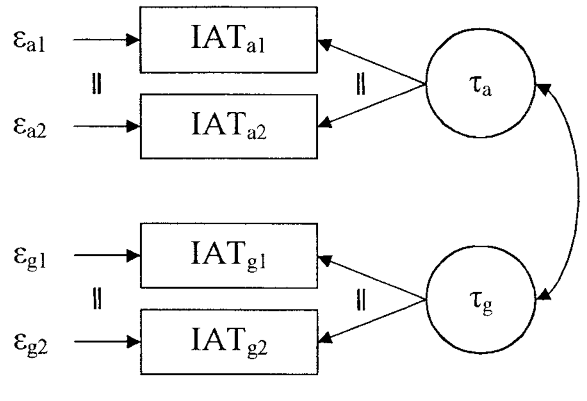

Figure 2. The structural equation model used to fit the data of Experiments 2 and 3. IAT$_{a1}$ and IAT$_{a2}$ refer to the observed first and second test half of the attitude IAT, respectively. IAT$_{g1}$ and IAT$_{g2}$ refer to the first and second test half of the geometry IAT, respectively. ε_{a1}, ε_{a2}, ε_{g1}, and ε_{g2} denote residual variance of the respective observed IAT measures. τ_a (attitude IAT) and τ_g (geometric IAT) capture the systematic variance within each IAT. Parameter restrictions are indicated by a vertical "=." IAT = Implicit Association Test.

ANOVA. A 2 (geometry vs. attitude IAT) × 2 (task switch vs. task repetition) × 2 (compatibility order) × 2 (mirrored stimulus–response mapping) mixed factorial ANOVA was performed on the data with repeated measures on the first two factors. This analysis revealed a main effect of task switching, $F(1, 63) = 207.10$, $p < .01$, that was due to larger IAT effects for task-switch than for task-repetition trials across both IAT measures. The mean effects were smaller for the attitude IAT than for the geometry IAT, $F(1, 63) = 86.99$, $p < .01$. Additionally, IAT effects, $F(1, 63) = 6.26$, $p < .05$, as well as the asymmetry in task-switching costs, $F(1, 63) = 11.24$, $p < .01$, were larger when the compatible phase came first. The mean IAT scores for task-switch and task-repetition trials can be found in Table 1.

Discussion

The results of Experiment 2 demonstrate reliable contamination of a flower–insect attitude IAT. As predicted, the geometrical and attitude IAT measures were correlated, even though there was no common content factor underlying the measures. The obtained correlation is significant, though not exceedingly high, and it is in the order of magnitude of correlations that have been reported to indicate validity of IAT measures. Although the geometry IAT does most likely not reflect preexisting associations, it definitely does not measure the same associations as the attitude IAT. Therefore, this correlation is a manifestation of reliable method-specific variance, probably introduced by interindividual differences in task-switching performance that affect both IAT measures irrespective of their content.

Concerning the quality of the IAT as a tool to assess interindividual differences, this finding indicates a rather unsatisfactory discriminant validity of the flower–insect attitude IAT inasmuch as the latter IAT also measures stable method-specific interindividual differences to a moderately high degree. It seems plausible that the high reliability estimates obtained in Experiments 1A, 1B, and 2 (and other IAT studies as well) partially reflect a common factor that does not represent interindividual differences in the construct to be measured.

Experiment 3

One obvious difference between the contingency-based IAT and the flower–insect IAT on the one hand and many IAT measures used in applied research on the other hand is that there are almost no interindividual differences in the direction of IAT effects in the former measures. In Experiment 2, for example, the IAT effects for almost all participants had the same direction, that is, most participants preferred flowers to insects, according to their IAT scores. Experiment 3 was conducted to test whether method-specific variance is also found in a self–other extraversion IAT with a pronounced variation in the direction of scores.

Table 2
Variances and Covariances of IAT Scores for Two Random IAT Test Halves in Experiments 2 and 3

IAT measure	IAT_{11}	IAT_{12}	IAT_{g1}	IAT_{g2}
Experiment 2				
IAT_{11}[b]	**17,233.25**	13,227.18	8,523.04	6,503.85
IAT_{12}[b]		**14,797.33**	6,859.61	4,904.25
IAT_{g1}[a]			**23,716.19**	18,878.98
IAT_{g2}[a]				**19,526.93**
Experiment 3				
IAT_{11}[c]	**21,463.23**	20,440.06	8,289.94	9,795.75
IAT_{12}[c]		**26,971.34**	8,964.79	11,946.67
IAT_{g1}[a]			**28,657.55**	25,187.76
IAT_{g2}[a]				**30,347.98**

Note. IAT_{11} and IAT_{12} refer to the random test halves of the attitude (Experiment 2) and extraversion (Experiment 3) IATs; IAT_{g1} and IAT_{g2} refer to the test halves of the contingency-based IATs. Boldface type indicates variance. IAT = Implicit Association Test.
[a] Contingency-based IAT with geometrical objects. [b] Flower–insect attitude IAT. [c] Self–other extraversion IAT.

Method

Each participant of Experiment 3 completed a self–other extraversion IAT (described below) followed by the contingency-based IAT with geometrical objects used in the previous experiments. Although we expect the extraversion IAT to be reliably contaminated by method-specific variance, it is important to note that this prediction refers to the absolute size of the extraversion IAT scores. Imagine, for example, 2 participants with reliably poor task-switching performance. Both participants will have highly positive scores in the contingency-based IAT. If the scores for the extraversion IAT are reliably contaminated, large effects are expected in this IAT as well, but they should be highly positive for extraverted and highly negative for introverted participants. Consequently, tests for method-specific variance refer to the absolute magnitude of the extraversion IAT scores.

Participants. Eighty-one (60 women and 21 men) student and nonstudent volunteers agreed to participate in return for detailed individual feedback and/or partial course credit. Their mean age was 26.14 years (SD = 8.57).

Material and questionnaire. Adjectives representing extraversion and introversion were chosen on the basis of typicality word norms provided by a study by Ostendorf (1994). It turned out that adjectives typical for extraversion were evaluated more positively than the introversion adjectives. As representing a concept with atypical, but evaluatively unconfounded, stimuli would not make sense, achieving a high typicality was nevertheless the most important criterion in material selection. Some extremely positive and negative adjectives could be removed from the pool without losing typicality, however. To prevent effects of word length, adjectives with more than 11 characters were removed. From the remaining adjectives, the 14 most typically extraverted and 14 most typically introverted adjectives were selected and used in the study. The self-versus-other dimension was represented by two sets of five words referring to the self (*I*, *self*, *me* [German: *mir*], *me* [German: *mich*], *own*) or to other people (*they*, *them*, *your*, *you*, *other*), taken from the study by Asendorpf et al. (2002). The stimulus words are listed in the Appendix.

The explicit measure used in Experiment 3 is the well-established German translation of the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1989, 1992), the so-called Neo-Fünf-Faktoren Inventar (Borkenau & Ostendorf, 1993). The NEO-FFI was used in the standard paper-and-pencil version.

Procedure. Each experimental session consisted of four major phases. In the first phase, participants completed the NEO-FFI (Borkenau & Ostendorf, 1993) to obtain an explicit measure of extraversion. The second phase consisted of an IAT in which a self-versus-other task and an extraversion-versus-introversion task were combined. For the extraversion-versus-introversion classification task, participants were instructed to classify the adjectives as either typically extraverted or typically introverted. Participants were given a definition of these terms that was based on the instructions used in the rating study by Ostendorf (1994). The order in which the extraversion + self and extraversion + other conditions were to be performed was counterbalanced. Following the extraversion IAT, participants performed the contingency-based IAT that was used in Experiments 1A and 2. Both IAT measures used three blocks for each combined phase and one block for each training phase. All other procedural parameters of the two measures remained identical to those of the previous experiments. In the fourth phase, participants were debriefed and received personalized feedback on their results in the questionnaire.

Results

Trials with incorrect responses (6.5%) were discarded from the IAT analyses as well as the first two trials of each block. Response latencies below 100 ms or above 3,000 ms (0.8%) were recoded to 100 ms and 3,000 ms, respectively.

The mean score for the NEO-FFI Extraversion scale was 2.47 (SD = 0.56), close to the mean of 2.36 (SD = 0.57) reported in the test manual (Borkenau & Ostendorf, 1993). Scores in the contingency-based IAT were computed by subtracting performance in the compatible condition from performance in the incompatible condition in all analyses. The score for the extraversion IAT was computed by subtracting performance in the self + extraversion condition from performance in the self + introversion condition for all analyses involving the explicit measure. Positive scores can be interpreted as a stronger self-extraversion association, and negative scores as a stronger self-introversion association. Mean IAT effects were found to be 369 ms (SD = 165) for the geometry material and −41 ms (SD = 205 ms) for the extraversion IAT. Scores in the extraversion IAT ranged between −768 ms and 446 ms. The median of the extraversion scores was −7 ms, indicating that approximately half of the sample had an

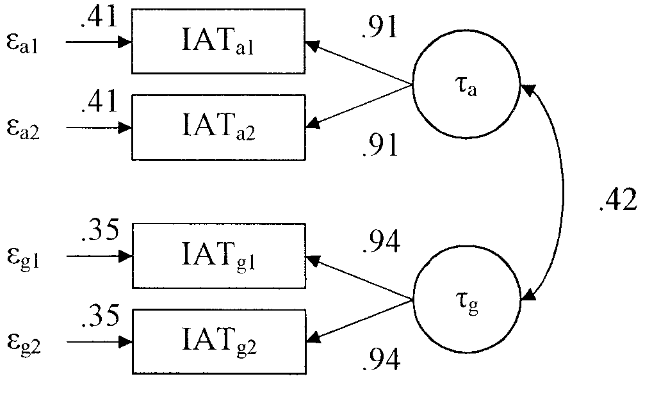

Figure 3. Standardized parameter estimates of the structural equation model in Experiment 2 for the attitude IAT (IAT_{a1} and IAT_{a2}) and the contingency-based IAT (IAT_{g1} and IAT_{g2}). IAT_{a1} and IAT_{a2} refer to the observed first and second test half of the attitude IAT, respectively. IAT_{g1} and IAT_{g2} refer to the first and second test half of the geometry IAT, respectively. ε_{a1}, ε_{a2}, ε_{g1}, and ε_{g2} denote residual variance of the observed IAT measures. τ_a (attitude IAT) and τ_g (geometric IAT) capture the systematic variance within each IAT. IAT = Implicit Association Test.

implicit self-concept of being more extraverted, and the other half of being more introverted.

Internal consistencies of the two IAT measures were again estimated by computing Cronbach's alphas between IAT scores based on four randomly assembled sets of trials for each of the two measures. The internal consistencies were $\alpha = .93$ for the contingency-based IAT, and $\alpha = .94$ for the extraversion IAT. Convergent validity of the extraversion IAT was estimated by correlating the signed extraversion IAT scores with the NEO-FFI Extraversion score. A small but significant correlation of $r = .24$ ($p < .05$) was obtained, indicating that implicit and explicit measures share a small amount of common variance. As expected, the correlation between the NEO-FFI Extraversion score and the geometry IAT was practically zero ($r = -.03$), and not significant.

To test for reliable contamination, the absolute magnitude of the extraversion IAT effect was correlated with the contingency-based IAT. This correlation was again significantly greater than zero ($r = .39$, $p < .01$), showing that participants with large effects in the IAT with geometrical objects tended to have larger effects in the extraversion measure as well. It should be noted that this does not reflect a difference between introverted and extraverted participants, but a correspondence between absolutely large and small effects in both IAT measures.

Structural equation analysis. Two test halves were randomly generated for the compatibility effects in the extraversion IAT (i.e., the size of the scores independent of their direction) and the contingency-based IAT. The covariance matrix between the resulting four IAT scores is given in Table 2. The structural equation model fit these data rather well, $\chi^2(5, N = 81) = 7.75$, $p = .17$. The correlation between the latent variables representing the extraversion and contingency-based IAT measures was .43 (SE = .10), and is significantly larger than zero, $t(80) = 4.23$, $p < .01$. The standardized parameter estimates obtained in this analysis are shown in Figure 4.

ANOVA. The IAT effects for the geometry and extraversion IAT measures were submitted to a 2 (geometry vs. extraversion IAT) × 2 (task switch vs. task repetition) × 2 (compatibility order) × 2 (reversed response mapping) mixed factorial ANOVA with repeated measures on the first two factors. This analysis revealed that the effects were smaller in the extraversion IAT than in the geometry IAT, $F(1, 77) = 140.15$, $p < .01$. The mean IAT effect was larger for task-switch than for task-repetition trials, $F(1, 77) = 162.32$, $p < .01$. These two factors were involved in a two-way interaction, showing that the difference between task-switch and task-repetition trials was larger in the geometry IAT, $F(1, 77) = 21.63$, $p < .01$. Two additional interactions indicate effects of compatibility order. Effects in the extraversion IAT were larger when the self + introversion condition was performed first, $F(1, 77) = 7.70$, $p < .01$, as was the difference between task-switch and task-repetition trials in this condition, reflected in a three-way interaction between compatibility order, task switching, and IAT type, $F(1, 77) = 22.77$, $p < .01$. The mean IAT scores for task-switch and task-repetition trials can be found in Table 1.

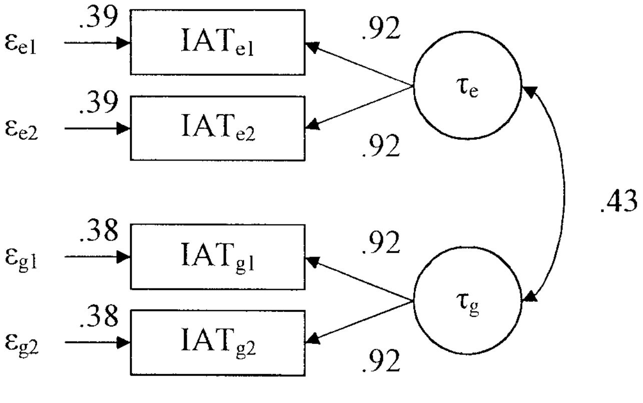

Figure 4. Standardized parameter estimates of the structural equation model in Experiment 3 for the extraversion IAT (IAT_{e1} and IAT_{e2}) and the contingency-based IAT (IAT_{g1} and IAT_{g2}). IAT_{e1} and IAT_{e2} refer to the observed first and second test half of the extraversion IAT, respectively. IAT_{g1} and IAT_{g2} refer to the first and second test half of the geometry IAT, respectively. ε_{e1}, ε_{e2}, ε_{g1}, and ε_{g2} denote residual variance of the observed IAT measures. τ_e (extraversion IAT) and τ_g (geometric IAT) capture the systematic variance within each IAT. IAT = Implicit Association Test.

Discussion

The main findings of Experiment 3 are a small but significant amount of shared variance between implicitly and explicitly measured extraversion, and a medium-sized, significant correlation between the absolute effect in the extraversion IAT and the contingency-based IAT measure. This combination of a small correlation between the implicit and explicit measures of extraversion, and a heterotrait-monomethod correlation between two reliable IAT measures is what is expected on the basis of the reliable contamination hypothesis. However, besides interindividual differences in task-switching costs, several other factors may have contributed to the low implicit–explicit correlation obtained in Experiment 3, such as balancing of compatibility order (Asendorpf et al., 2002), and differences in the constructs tapped by implicit and explicit measures (Bosson et al., 2000; Wilson, Lindsey, & Schooler, 2000). Additionally, procedural variations like a relatively long response–stimulus interval of 800 ms and the incorporation of task-repetition trials may have played a role. However, because task-switching variance in IAT scores was observed to be larger with short response–stimulus intervals (Mierke & Klauer, 2001), and because the resulting score should be more strongly affected by task-switching when no repetition trials are used, these factors may in fact lead to an underestimation of the impact of task-switching variance.

Before discussing implications of the findings in the General Discussion section, several methods for reducing the impact of reliable contamination, namely removing task-switch trials from the analysis, using different scoring procedures, and partialling out method-specific variance, will be evaluated in the next section.

Reducing Method-Specific Variance

As already mentioned, reliable contamination by method-specific variance has a number of psychometric consequences: (a) It directly conflates estimates of retest reliability and internal consistency; (b) it conflates correlations between the absolute scores of any two IATs, even if they are not related by shared content; and (c) it reduces the convergent validity of the IAT as elaborated below. The problem of method-specific variance thereby contributes to explaining the known discrepancy between high internal consistency and moderate validity of IAT measures, but it does not challenge the existing evidence on the predictive validity of the IAT for criteria that are based on other methods (such as self-report measures). To the contrary, reducing method-specific variance in IAT scores should even increase correlations with such criteria.

To see this, consider how the correlation between the implicit and the explicit measure of extraversion in Experiment 3 is affected by partialling out the indicator of method-specific variance, that is, the geometry IAT, from the extraversion IAT. Assuming that the correlation between the explicit extraversion measure and the geometry IAT measure is zero, the covariance between the implicit and explicit extraversion measures should not be affected by partialling out the geometry IAT scores. The correlation between the extraversion measures, however, should be increased because of a smaller amount of error variance in the corrected implicit measure. The partial correlation can then be estimated by a simple equation:

$$r_{IE.G} = \frac{\mathrm{COV}_{IE}}{\sqrt{(1 - r_{IG}^{2})\ \mathrm{var}_{I}\ \mathrm{var}_{E}}} = \frac{1}{\sqrt{1 - r_{IG}^{2}}}\, r_{IE},$$

where *I* is the extraversion IAT score, *G* is the geometry IAT, and *E* is the explicit extraversion score. If an indicator of method-specific variance is known, a factor can be computed reflecting the increase in magnitude for a correlation of the corrected IAT scores with an explicit measure when method-specific variance is controlled for. However, because sign differences in the extraversion IAT scores will conceal the amount of method-specific variance, the correlation r_{IG} is not a suitable measure of method-specific variance in the case of Experiment 3. Therefore the correlation between the geometry IAT and the absolute extraversion IAT scores obtained in Experiment 3 was inserted into Equation 1. Correlations between corrected scores and the explicit measure are estimated to be 1.0859 times larger than the correlations originally obtained. Although this constitutes only a moderate increase, it may make a considerable difference in terms of test power for detecting nonzero correlations when they exist.

To evaluate this analysis empirically, we regressed the absolute values of the extraversion IAT scores on the scores in the geometry IAT. Method-specific variance should be removed from the regression residuals. To reassign the sign information contained in the original IAT scores, the regression residuals were shifted so that the smallest residual was zero, and multiplied by plus or minus one, depending on the sign of the original effect in the extraversion IAT. Shifting the residuals was necessary to ensure the integrity of the sign information, that is, to reintroduce the vertical separation of positive and negative IAT scores that is removed in the regression analysis of absolute extraversion IAT scores. With a large effect in the geometry IAT, for example, the residual of an originally small, but positive effect in the extraversion IAT would otherwise be negative, because of a large divergence from the predicted value. The correlation between these corrected scores and the explicit measure turned out to be higher—to the expected small degree—than the one originally obtained ($r = .27$, $p < .01$).

Method-specific variance may also be reduced by using alternative algorithms to compute the IAT scores (Greenwald, Nosek, & Banaji, 2003). One advantage of this approach is that no specific indicator of method-specific variance is required. The effects of three simple algorithms were to be explored, namely (a) a simple scoring algorithm based on untransformed response latencies with latencies below 300 ms recoded to 300 ms and latencies above 3,000 ms to 3,000 ms with the first two trials of each block discarded ($\mathrm{IAT}_{\mathrm{RT}}$), (b) the scoring algorithm suggested by Greenwald et al. (1998) based on an additional logarithmic transformation of the latencies ($\mathrm{IAT}_{\mathrm{LN}}$), and (c) the D measures (see Greenwald et al., 2003, for details on the complete algorithm), in which the IAT effects are based on untransformed response latencies and scaled in units of the individuals' standard deviations. Because the results for the different variations of the latter measure were almost identical, they are only reported for the first variation, which incorporates neither a lower tail treatment nor an error penalty ($\mathrm{IAT}_{\mathrm{D1}}$).

Another, theoretically motivated approach to reducing method-specific variance without incorporating an additional indicator of method-specific variance can be derived from the task-switching account. The model suggests that data from task-repetition trials should be less affected by interindividual differences in task-switching costs than data based on task-switch trials. Although task-switching costs are believed to affect several successive trials and are therefore unlikely to be completely eliminated after only one task repetition (e.g., Allport et al., 1994), their impact on task-repetition trials is expected to be smaller than on task-switch trials. This suggests computing IAT effects on the basis of only the task-repetition trials, and leads to the predictions that (a) task-repetition IAT scores should contain a smaller amount of method-specific variance than IAT scores based on all trials or, a fortiori, IAT scores based on task-switch trials, and (b) the task-repetition IAT scores should simultaneously have higher convergent validity (see above).

To test these predictions as well as the impact of the scoring procedures, the flower–insect IAT and the extraversion IAT were scored with the algorithms described above and correlated with the geometry IAT as an indicator of method-specific variance. Because the amount of method-specific variance should be maximal in this indicator, only task-switch trials were included, and scored on the basis of untransformed response latencies as described in the results sections of Experiments 1–3. To assess convergent validity, the explicit measure of extraversion was the criterion to be predicted.

The results of these analyses are shown in Table 3. The most prominent result is that the amount of method-specific variance in the IAT scores can be dramatically reduced by using the new scoring procedures. Method-specific variance was completely removed from the flower–insect IAT of Experiment 2, and markedly reduced in the extraversion IAT. Furthermore, there was no effect of using only task-switch or only task-repetition trials on method-specific variance in the IAT scores computed with the new D measures, which is additional evidence for the conclusion that method-specific variance was removed from these scores.

The correlation indicating method-specific variance in the log-transformed and untransformed IAT scores was significant when task-switch trials were analyzed. As predicted, a considerably lower amount of method-specific variance was obtained when only task-repetition trials were analyzed. The difference between the correlations obtained with task-switch trials and the correlations obtained with task-repetition trials was significant for both scoring methods in Experiment 2 ($z = 1.66$, $p < .05$ for $\mathrm{IAT}_{\mathrm{RT}}$; $z = 1.81$,

Table 3
Correlations of Differently Scored Substantial IATs With a Geometry IAT (Method-Specific Variance) and With an Extraversion Scale (Construct-Specific Variance; Experiment 3)

		Method-specific variance		
Type	Measure	Exp. 2	Exp. 3	Construct-specific variance: Exp. 3
Task-switch trials	IAT_{RT}	.40**	.37**	.21
	IAT_{LN}	.31*	.29**	.26*
	IAT_{DI}	.08	.12	.25*
All trials	IAT_{RT}	.33**	.34**	.26*
	IAT_{LN}	.21	.26*	.30**
	IAT_{DI}	−.02	.16	.28*
Task-repetition trials	IAT_{RT}	.13	.23*	.32**
	IAT_{LN}	.00	.18	.34**
	IAT_{DI}	−.12	.06	.28*

Note. IAT_{RT} refers to scores obtained with untransformed response latencies; IAT_{LN} refers to scores obtained with log-transformed response latencies; IAT_{DI} refers to the first variation of the new D measures proposed by Greenwald et al. (2003; see text). Results for the scores based on task-switch trials provide the best approximation to the procedures originally proposed by Greenwald et al. (1998). IAT = Implicit Association Test; Exp. = Experiment.
* $p < .05$. ** $p < .01$.

$p < .05$ for IAT_{LN}) but not significant in Experiment 3, although the correlations point in the expected direction. It should be noted that the scores based on task-switch trials provide the best approximation and basis of comparison with the procedures originally proposed by Greenwald et al. (1998) and used in many subsequent studies, in which task switching is required on every trial. A significant difference between the correlations indicating method-specific variance in the D measures based on task-switch trials and the correlations obtained with the untransformed scores ($z = 1.94$, $p < .05$ for Experiment 2; $z = 1.67$, $p < .05$ for Experiment 3) suggests that method-specific variance was indeed reduced by the new algorithm.

The correlation with the explicit measure of extraversion was more or less the same for each scoring procedure, but descriptively higher correlations were obtained with the scores based on task-repetition trials. The highest correlation with the explicit measure was obtained for the log-transformed measure based on task-repetition trials. This relatively better performance of the task-repetition IAT scores emerges despite the fact that analyzing only task-repetition trials has the consequence of diminishing the number of items/trials by a factor of two, and that the mean IAT scores were generally lower for task-repetition trials in all experiments reported here.

General Discussion

The results of the four experiments reported in this article lead to a number of conclusions concerning the IAT procedure as a measure of interindividual differences. Most importantly, the results show that conventionally scored IAT effects contain both stable content-specific and stable method-specific variance. Method-specific variance emerged in the form of correlations between IAT measures that were designed to have no overlap with respect to content. It could be reduced by computing IAT scores on the basis of only task-repetition trials, and even more effectively by using the D measures that scale IAT effects in units of standard deviations (Greenwald et al., 2003). Construct-specific variance estimated by the correlation with the explicit measure of extraversion was more or less independent of these operations, but slightly higher for scores based on task-repetition trials.

Practical Consequences of Method-Specific Variance

As already mentioned, there are a number of practical problems associated with reliable method-specific variance that should be considered, especially when interpreting results obtained with the standard scoring procedures. One of these problems is that method-specific variance can directly conflate internal consistencies and retest reliabilities, which may explain why reliabilities of IAT measures are sometimes found to be high, whereas validity coefficients are poor (e.g., Bosson et al., 2000; Farnham, 1999; Greenwald & Farnham, 2000; Karpinski & Hilton, 2001; Ottaway et al., 2001; Rudman et al., 1999). Thus, differences between IAT effects of the same sign, though reliable and consistent, may be less informative with respect to interindividual differences in the constructs to be measured than expected. This is particularly problematic for the interpretation of test scores at the individual level. For a single participant, method- and content-specific variance cannot be disentangled by repeated measurement, or an increased number of trials, as they are unavoidably confounded. In experimental situations, however, content-unrelated interindividual differences are cancelled out by random assignment to the experimental groups and, thus, the internal validity of experiments would not be threatened. However, external validity may be an issue when an experimental manipulation is likely to affect task-switching performance. For example, a mood manipulation might affect the amount of available cognitive resources and thereby, the capability to perform task switches. If so, differences in mean scores of a self-esteem IAT between the experimental conditions need not reflect effects of mood on implicit self-esteem, but might be due to effects on task-switching performance. Similarly, in quasi-experimental settings, group membership may be confounded with factors known to be related to task-switching performance. Comparing implicit self-esteem of young and elderly participants, for example, may result in group differences that are overestimated because of reliable contamination.

Implications for Theoretical Accounts of the IAT

The surprisingly good performance of the new scoring algorithm based on scaling the IAT effects in units of the individuals' standard deviations (Greenwald et al., 2003) indicates that method-specific variance is related to the overall variability of response latencies. One possibility is that method-specific variance resides in the fastest and slowest latencies rather than in the average latencies. Locating the causes of method-specific variance in the tails of the latency distribution is compatible with the assumption that extreme latencies can occur when an additional cost for switching between task sets affects responding in the incompatible condition. The relation of task-switching costs and method-specific variance is supported by the finding of greater proportions of method-specific variance in task-switch than in task-repetition trials. However, future research will have to address more explicitly whether reliable contamination is specifically caused by task-switching costs or by a more general performance factor such as cognitive speed or working-memory capacity.

As yet, the general pattern of findings confirms the predictions derived from the task-switching account, but it does not conclusively rule out the alternative accounts by De Houwer (2001) and Brendl et al. (2001) presented earlier. However, these alternative accounts cannot explain the presence of method-specific variance without additional assumptions. Brendl et al., for example, explicitly assume that relevant and irrelevant stimulus–response mappings operate in parallel and lead to a lower net accumulation rate in the random-walk process, when they are mapped incompatibly. Switching between task sets must be conceptualized as a process of adapting weights for relevant and irrelevant information before—or in addition to—an accumulation of response-related information on a decision axis, in the theoretical framework of a random-walk model; but these necessary modifications of the account by Brendl et al. have not been specified so far. Unlike the models mentioned above, the figure-ground asymmetry model by Rothermund and Wentura (2001) predicts method-specific variance without further assumptions. However, the results of Experiment 1A are difficult to explain with this model. We thus believe that asymmetries in salience, even though they may be sufficient to produce IAT effects, are not a necessary precondition. Rather, these asymmetries appear to be a subset of a more general set of relations that can produce IAT effects.

Implications for Applied IAT Research

Another implication of the findings reported here is that the set of relations producing IAT effects is not restricted to preexisting associations. Other types of relation, like a contingency between visual properties of the stimuli, can cause effects even when the response categories are completely unrelated, as in the case of size and color of objects. Such relations based on superficial stimulus properties can be problematic for substantial IAT measures as well, for example, when the relation causing the IAT effects is based on unintended confoundings in the stimulus material, such as systematic differences in familiarity or salience instead of the intended differences in the strength of associative relations. The results reported here suggest that measures based on such unintended relations may even exhibit encouraging psychometric properties in terms of internal consistency and retest or parallel-forms reliability due to stable method-specific variance, at least when standard scoring methods are used.

One should note that similar caveats hold for explicit self-report measures, which have long been known to contain variance components that are not related to the construct to be measured. A number of response styles such as acquiescence styles and differential tendencies toward socially desirable responses have been identified, and methods to control for them have been discussed (Wilde, 1977). It should therefore not come as a surprise that IAT measures are likewise contaminated by a variance component that is not related to the purpose of the measurement. The contribution of the present article is to demonstrate that such method-specific variance exists, to assess its nature and size as well as its contribution to measures of reliability and validity, and finally, to propose methods to control for it. Among these methods, the scoring algorithms recently suggested by Greenwald et al. (2003) produced the most convincing results. Even though there is currently no clear-cut account for how the new algorithms work, the convincing results obtained for the data reported in this article suggest that the new scoring procedures are superior to the conventional algorithms and should be used in future research with the IAT, either instead of or in conjunction with the standard scoring procedures.

References

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

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83,* 380–393.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger III & J. S. Nairne (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder. Science conference series* (pp. 117–150). Washington, DC: American Psychological Association.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift fuer Experimentelle Psychologie, 48,* 145–160.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI)—Handanweisung* [NEO Five-Factor Inventory (NEO-FFI)—Manual]. Göttingen, Germany: Hogrefe & Huber.

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631–643.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 760–773.

Celex. (1995). *The Celex lexical database, Release 2.* Nijmegen, the Netherlands: Center for Lexical Information.

Costa, P. T., & McCrae, R. R. (1989). *The NEO-PI/FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory. Professional manual.* Odessa, FL: Psychological Assessment Resources.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 121,* 163–170.

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. (2001). A structural and process analysis of the implicit association test. *Journal of Experimental Social Psychology, 37,* 443–451.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology, 83,* 1441–1455.

Farnham, S. D. (1999). *From implicit self-esteem to in-group favoritism.* Seattle: University of Washington.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79,* 1022–1038.

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. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift fuer Experimentelle Psychologie, 48,* 85–93.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: 1. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774–788.

Kramer, A. F., Hahn, S., & Gopher, D. (1999). Task coordination and aging: Explorations of executive control processes in the task switching paradigm. *Acta Psychologica, 101,* 339–378.

Kray, J., & Lindenberger, U. (2000). Adult age differences in task switching. *Psychology and Aging, 15,* 126–147.

Maison, D., Greenwald, A. G., & Bruin, R. (2001). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin, 2,* 61–79.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

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

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive Psychology, 41,* 211–253.

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

Monsell, S., Yeung, N., & Azuma, R. (2000). Reconfiguration of task-set: Is it easier to switch to the weaker task? *Psychological Research/Psychologische Forschung, 63,* 250–264.

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

Ostendorf, F. (1994). Zur Taxonomie deutscher Dispositionsbegriffe [A taxonomy of German dispositional concepts]. In W. Hager & M. Hasselhorn (Eds.), *Handbuch deutschsprachiger Wortnormen* [Handbook of German word nouns] (pp. 382–441). Göttingen, Germany: Hogrefe & Huber.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition, 19,* 97–144.

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

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift fuer Experimentelle Psychologie, 48,* 94–106.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition, 17,* 437–465.

Sohn, M. H., & Anderson, J. R. (2001). Task preparation and task repetition: Two-component model of task switching. *Journal of Experimental Psychology: General, 130,* 764–778.

Sohn, M. H., & Carlson, R. A. (2000). Effects of repetition and foreknowledge in task-set reconfiguration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1445–1460.

Swanson, J. E., Rudman, L. A., & Greenwald, A. G. (2001). Using the Implicit Association Test to investigate attitude-behaviour consistency for stigmatised behaviour. *Cognition and Emotion, 15,* 207–230.

Wilde, G. J. S. (1977). Trait description and measurement by personality questionnaires. In R. B. Cattell & R. M. Dreger (Eds.), *Handbook of modern personality theory* (pp. 69–103). Washington: Hemisphere Publication Services.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101–126.

Wylie, G., & Allport, A. (2000). Task switching and the measurement of "switch costs." *Psychological Research/Psychologische Forschung, 63,* 212–233.

(*Appendix follows*)

Appendix

Stimulus Words Used in the Extraversion Implicit Association Test [and English Translations in Brackets]

Self: *ich* [I], *selbst* [self], *mir* [me], *mich* [me], *eigen* [own].

Other: *sie* [they], *ihnen* [them], *euer* [your], *ihr* [you], *andere* [others].

Extraversion: *wagemutig* [adventurous], *personenorientiert* [sociable], *vergnügungsfreudig* [hedonistic], *ungehemmt* [unrestrained], *draufgängerisch* [reckless], *redselig* [gabby], *dominierend* [dominant], *leidenschaftlich* [passionate], *energiegeladen* [energetic], *expressiv* [expressive], *mitteilsam* [communicative], *dominant* [dominant], *geschwätzig* [loquacious], *gesprächig* [talkative].

Introversion: *zurückhaltend* [contained], *schüchtern* [shy], *zurückgezogen* [withdrawn], *still* [quiet], *distanziert* [distant], *einzelgängerisch* [aloof], *reserviert* [reserved], *unauffällig* [inconspicuous], *scheu* [bashful], *schweigsam* [taciturn], *unspontan* [non-spontaneous], *verschwiegen* [discreet], *vorsichtig* [cautious], *ruhig* [calm].

Received July 24, 2002
Revision received March 7, 2003
Accepted April 8, 2003 ■