Implicit Association Test: Separating Transsituationally Stable and Variable Components of Attitudes toward Gay Men

Melanie C. Steffens[1] and Axel Buchner[2]

[1] University of Trier, Germany and [2] Heinrich-Heine University, Düsseldorf, Germany

Abstract. Implicit attitudes are conceived of as formed in childhood, suggesting extreme stability. At the same time, it has been shown that implicit attitudes are influenced by situational factors, suggesting variability by the moment. In the present article, using structural equation modeling, we decomposed implicit attitudes towards gay men into a person factor and a situational factor. The *Implicit Association Test* (Greenwald, McGhee, & Schwartz, 1998), introduced as an instrument with which individual differences in implicit attitudes can be measured, was used. Measurement was repeated after one week (Experiment 1) or immediately (Experiment 2). Explicit attitudes towards gay men as assessed by way of questionnaires were positive and stable across situations. Implicit attitudes were relatively negative instead. Internal consistency of the implicit attitude assessment was exemplary. However, the within-situation consistency was accompanied by considerable unexplained between-situation variability. Consequently, it may not be adequate to interpret an individual implicit attitude measured at a given point in time as a person-related, trait-like factor.

Key words: implicit attitudes, reliability of measurement, Implicit Association Test, attitudes toward gay men

Self-reports have been the tool for the attitude researcher for decades now – not so much the tool of choice, though, given a long list of well-known criticisms (e.g., Nisbett & Wilson, 1977). Rather, it has been the tool used because there was no alternative. The problems of self-report data hit home all the more with socially sensitive issues, for instance, attitudes towards gay men. Measuring implicit attitudes instead is a potential solution that has recently become very prevalent, especially using the Implicit Association Test (IAT, Greenwald et al., 1998). On the one hand, these implicit attitudes have been considered determined early in life and resistant to change, even in the face of consciously endorsed divergent attitudes (Devine, 1989; Wilson, Lindsey, & Schooler, 2000), suggesting to the poor human that even if she is willing to act and think in an egalitarian way, the early-learned and deeply-rooted immediate and uncontrollable reactions originating in her "subconscious" tell a different story. On the other hand, researchers have shown that implicit attitudes and stereotypes, in the hands of the expert, can be changed with apparent ease by some simple preceding task or situational factor, including such subtleties as the race of the experimenter (Lowery, Hardin, & Sinclair, 2001). Little, however, is known about the stability or variability of implicit attitudes if they are not manipulated. This is the question we address in the present article, using the IAT in combination with confirmatory factor-analytic models to assess attitudes towards gay men.

The article was written while the first author stayed at Yale University, supported by grant Ste 938/3–1 from the Deutsche Forschungsgemeinschaft. Parts of this research were presented at the *Workshop Implizite Diagnostik* in Heidelberg, January 2000, sponsored by the Deutsche Forschungsgemeinschaft, and at the 109th Annual Convention of the American Psychological Association, San Francisco, CA, USA, 24–28 August, 2001. For help and suggestions in various stages of this research project, we would like to thank Roland Neumann as well as Steve Arendt, Alexander Besemer, Pascale David, Mirjam Halder, Melanie Loch, Nicola Meyer, Katrin Modabber, Stefanie Peters, and Fabian Schüßler.

DOI: 10.1027//1618-3169.50.1.33

The Implicit Association Test (IAT)

The IAT's rationale is that people are able to react fast if a pair of closely associated categories requires one reaction and another pair, another reaction. In this case, the category–response assignment is "congruent." In contrast, if closely associated categories require different reactions so that the category–response assignment is incongruent, reactions should be relatively slow. The difference in reaction times between the incongruent and the congruent task, called the IAT effect, is taken to be an indicator of the association between the categories used. If one of those categories is an *evaluative category* (e.g., words are to be judged as "positive" vs. "negative"), then the IAT effect may be an indicator of a person's attitude towards the *target category* (e.g., "gay men" vs. "heterosexuals").

Consider a person's average reaction time in a task in which (a) words closely related to the "heterosexual" concept (e.g., wedding) and words with a clearly "positive" valence (e.g., good) require the same reaction, and (b) words associated with "gay" and with "negative" require a different reaction (henceforth, the heterosexual + positive task). This reaction time is compared to a task in which associates of "gay" and "positive" require the one response, and associates of "heterosexual" and "negative," the other (henceforth, the gay + positive task). If people react faster in the heterosexual + positive than in the gay + positive task, then "heterosexual" and "positive" seem to be more closely associated for them than "gay" and "positive": Their implicit attitude toward "heterosexual" seems more positive than towards "gay."

Given the large effect sizes observed in IATs, one may hope that IATs can be used for "measuring individual differences in implicit cognition" (Greenwald et al., 1998, p. 1464). This would only be the case if IATs had psychometric qualities that allowed for individual diagnosis. If these qualities could be established, then numerous applications of the procedure are conceivable that may fundamentally change the field of psychological assessment: For instance, which of the managers in company X need to be sent to gender training? Which of the teachers in school Y show unacknowledged negativity towards children of different ethnicities? A precondition for such a conception of implicit measures is, however, that there is a large amount of transsituational stability in the implicit attitudes measured; that is, a factor should emerge that is sometimes referred to as "person factor" and that is, statistically, closely related to the reliability of measurement. Surprising as it may seem, however, not much work establishing IATs' psychometric qualities has been undertaken yet.

It is clear enough by now, though, that IATs can be used to measure group differences in implicit cognition. An IAT evaluating Japanese versus Koreans discriminated almost perfectly between Japanese and Korean test takers (Greenwald et al., 1998). Similar results were shown for the evaluation of Jewish versus Christian (Rudman, Greenwald, Mellott, & Schwartz, 1999). A gender stereotyping IAT predicted social competence ratings of stereotypically male acting candidates in a job interview situation (Rudman & Glick, 2001; Steffens, Günster, & Mehl, 2001b). The implicit association of self + conscientious predicted the number of errors made in a concentration test taken without time limits (Steffens,

Table 1. IAT Test-Retest Correlations Found in Previous Experiments.

Study	Kind of IAT and details	$r_{\tau\tau 2}$
Banse et al. (2001)	Attitudes toward homosexuality; variation of procedural factors between IATs, Experiment 1	.59
	Experiment 2; homogeneous sample (heterosexual males)	.38
Bosson et al. (2000)	Self esteem	.69
Cunningham et al. (2001)	Attitudes towards black people and white people; 1-week interval; correlations among latent variables were much higher	.31
Dasgupta et al. (2000)	Racial attitudes, name IAT versus picture IAT	.39
Dasgupta & Greenwald (2001)	Experiment 1, racial attitudes; correlation summed over groups after successful attitude manipulation	.65
Greenwald et al. (1998)	Experiment 2, attitudes towards Koreans versus Japanese; extreme groups (Korean and Japanese participants); full versus truncated Japanese names	.85
	Experiment 3, racial attitudes; male versus female names	.46
Greenwald & Farnham (2000)	Self esteem, Experiment 1, self-affect versus self-evaluation	.43
	Experiment 2, idiographic versus generic items	.68
	Additional data, generic items, varying delays	.52
Steffens (2002a)	Self-extraversion association	.61

Note. Unless mentioned, correlations were obtained with an immediate retest.

2002a). Finally, an IAT was the only implicit-self-esteem measure of seven such measures that correlated significantly with several criterion variables (Bosson, Swann, & Pennebaker, 2000).

It is a different question, however, whether a test that shows expected group differences measures reliably enough to allow for individual diagnosis. Whereas the reliability of many instruments used in implicit social cognition research has not yet been investigated thoroughly, some implicit memory tests which are similar to those instruments have been found wanting with regard to their reliability (Buchner & Brandt, in press; Buchner & Wippich, 2000; Meier & Perrig, 2000). We deem the assessment of IATs' reliabilities much more important than that of other measures used in implicit social cognition because these other measures typically show such small effect sizes that individual diagnosis is out of the question in the first place. When evaluating IATs' psychometric qualities, it is important to keep in mind that the IAT is not a single standardized research instrument, but a whole family of tests that do not necessarily have more in common than Likert-type scales assessing different subjects, so that it may be "necessary to evaluate the psychometric properties of any new implementation of the IAT" (Banse, Seise, & Zerbes, 2001, p. 146).

The little data that are available on IATs' reliability are presented in Table 1 (also see Dovidio, Kawakami, & Beach, in press; Greenwald & Nosek, 2001). Given these diverse test-retest correlations, individual diagnosis based on implicit attitudes seems rather unreliable. Many factors influencing reliability have not been investigated yet (e.g., length of IAT, stimuli). In one of the few studies that were directly aimed at testing an IAT's reliability, Cunningham, Preacher, and Banaji (2001) arrived at the conclusion that, with appropriate measurement models, the IAT could be regarded as a reliable measurement instrument. Using structural equation modeling techniques that are not described in much detail in their very concise article, they showed that even the lowest IAT test-retest correlations ever reported (as low as $r = .16$) contain a sizeable stable component, in addition to a large measurement error. As far as we can tell from the descriptions in their article, these authors decomposed the IAT variance into two components only, namely, a transsituational and an error component, thus ignoring the possibility that the IAT, at any given occasion, may additionally measure a consistent situational factor. Indeed, recent research manipulating situational factors has shown that the IAT is sensitive to situational variations. West Germans show a stronger implicit preference for this ingroup if ingroup identification was primed before taking the IAT (Kühnen, Schiessl, Bauer, Paulig, Pöhlmann, & Schmidthals, 2001). Implicit attitudes towards black people are less negative if positive black and negative white figures have been reviewed prior to attitude assessment (Dasgupta & Greenwald, 2001). Implicit attitudes towards the elderly, as measured in the IAT, were more positive if a good + elderly association was practiced before taking the IAT (Karpinski & Hilton, 2001). Similarly, the implicit stereotype of women as weak was reduced if a strong woman had been imagined in a mental imagery task before the IAT was taken (Blair, Ma, & Lenton, 2001). What if no attempt at manipulating the situation is made? We investigated whether the IAT measures a random person-situation interaction, too.

Confirmatory Factor-Analytic Models Assessing Stability and Change

Standard assessments of reliability have been criticized on various grounds (e.g., Bohrnstedt, 1993). Model-based reliability analyses are clearly preferable. Our approach is similar to that of Cunningham et al. (2001) in being based on latent variables. Specifically, we used a family of models introduced and discussed by Steyer (1989; see also Steyer, Majcen, Schwenkmezger, & Buchner, 1989). Essentially, these models are structural equation models that implement certain theoretical assumptions in terms of testable model restrictions. The minimal data structure that is needed to apply these models consists of two measurements of the same property at each of two measurement occasions. Within each occasion, the two measurements may be obtained by the common odd–even split of a test into two test halves, resulting in four measurements. Three different models were fit to these four measurements in the experiments described further on, the *reliability model*, the *stability model*, and the *consistency model*. The classical reliability model (see upper panel of Figure 1) assumes that the four measurements can be conceptualized as equivalent measurements of one single underlying true-score variable, τ. Similar to the approach taken by Cunningham et al. (2001), all variance not accounted for by the true-score variable τ is error variance (denoted by ε_{ij} in Figure 1, where i represents the measurement occasion and j represents the test half within a measurement occasion). The reliability model assumes that there is no situation-specific variance in measurement; that is, all variance that is not error variance is due to the transsituational factor. If these assumptions hold, the reliability model will fit the data. We used the χ^2 statistic and the root mean square error of approximation (RMSEA) to evaluate model fit, which is considered good if the RMSEA $< .05$ (see Bollen & Long, 1993; Kaplan, 2000).

The stability model (center panel of Figure 1) assumes that two different, but correlated true-score variables τ_1 and τ_2 generated the data at the first and second occasion, respectively. If this model fits the data, then the most interesting parameter is the correlation between the true-score variables, which may be obtained from the standardized solution when fitting the model to data. The higher this correlation, the more reliable is the measurement over time, or the more stable is the construct being measured.

For our situation of two measurements at each of two measurement occasions, the consistency (lower panel of Figure 1) and stability models are data equivalent in that both models imply the same theoretical variance-covariance matrix. However, each model has the advantage of delivering information that the other obscures. Whereas the correlation between the latent variables is shown only in the stability model, the consistency model allows us to decompose the true-score variance at each measurement occasion into two additive components. The first, situational component represents the variance that is specific to the measurement occasion, ζ. The second component is the variance that is common to the two true-score variables τ_1 and τ_2, ξ. This second component is a stable, transsituational component, the person effect. This model puts us in the position to compare how much of the true-score variable is accounted for by situational effects and how much is stable across measurement occasions. That is, in addition to the error variance estimated, a variance component is estimated that is not random noise, but situation-specific (or a situation–person interaction), and a third component that is stable across situations. Obviously, the situational component should be as small as possible *for a good instrument that measures a stable construct*. Note that apart from this pragmatic advantage, the consistency model may be regarded preferable to the stability model because it assumes explicitly that the same underlying latent variable (in the present case: a stable implicit attitude) is being measured at the two occasions.

Attitudes Towards Gay Men

Fernald (1995) gives a comprehensive review of attitudes towards, stereotypes of, and behavior towards gay men. Both the correlates of negative attitudes towards gay men and many cultural and individual determinants of anti-gay attitudes and behaviors are well-known. Analyses of the polls over the last twenty years show that attitudes towards gay men are getting less negative in industrialized countries such as the USA and Germany (see Steffens & Wagner, 2002, for a review). This trend is reflected in questionnaire findings assessing attitudes towards homo-

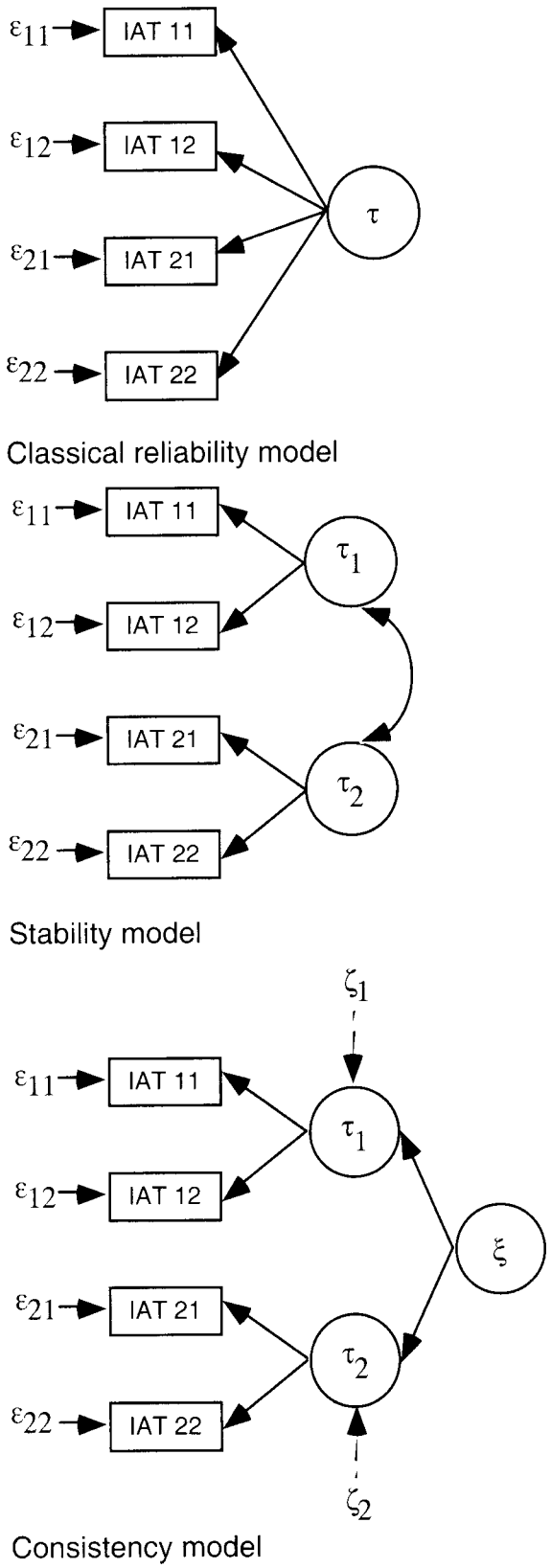

Figure 1. Three structural equation models which were fitted to the data of Experiments 1 and 2. Measured variables are displayed in rectangles, latent constructs, in circles (see text for details).

sexuals or towards gay men or lesbians (e.g., Herek, 1994). However, *implicit* attitudes might not be that positive. As of this writing, there appears to be only one published study looking at attitudes towards homosexuals in general (Banse et al., 2001). In the following experiments, we assessed attitudes towards gay men both implicitly using an IAT and explicitly using questionnaires. The two measurement occasions were one week apart in Experiment 1 but only 10 minutes in Experiment 2.

Experiment 1

We expected to find implicit negativity towards gay men (i.e., an IAT effect) in that participants react faster in the heterosexual + positive task than in the gay + positive task. This effect should be replicated a week later. Implicit and explicit assessments of attitudes should be reliably replicated after a week.

Method

Participants

Of the 103 students of the University of Trier, who participated for course credit at the first measurement occasion, a total of 84 (19 male) returned one week later. They were not informed about the topic of the experiment beforehand. Their mean age was 23 years (SD = 3.6). According to a Kinsey scale ranging from 1 (exclusively heterosexual) to 7 (exclusively homosexual), about 20% of them were not heterosexual; that is, they checked values of 3 or more. We included these participants in the sample in order not to reduce variance and thus provoke lowered estimates of reliability.

Materials

Two sets of stimuli were needed as IAT items, words for the target category and words for the evaluative category. For the evaluative category, adjectives with distinctly positive and negative valence were selected from German word norms (Hager & Hasselhorn, 1994). Words were selected such that they had no obvious relation to the concepts "heterosexual" or "gay" (see Steffens, Banaji, Jelenec, Wender, Anheuser, Goergens, Hülsebusch, Lichau, & Still, 2001a; Steffens & Plewe, 2001). The length of adjectives was between four and six letters. On a scale from −20 to 20 the average rating of the negative and positive adjectives was −12 and 15, respectively.

Pairs of names were used as instances of the target category in order to facilitate unambiguous associations. Name pairs were introduced as couples, two male names for gay couples (Christian + Felix; Lukas + Mark; Thomas + Philip; Daniel + Lars; Jörg + Erik) and a male and a female name for heterosexual couples (Michael + Sarah; Laura + Paul; Jochen + Sophie; Julia + Sven; Nils + Lisa). All names were very common and typical of 20 to 40 year olds in Germany. For each female name in the heterosexual couples' list, a male name was selected for the gay couples' list that was parallel with respect to the length (in terms of syllables) and associated ethnicity (e.g., having a "Northern" connotation). The rest of the male names were also pairwise parallel regarding these criteria, and it was randomized which pair member was assigned to the gay couples' list.

We developed an ad-hoc explicit attitude questionnaire that consisted of 28 statements. Ten of these were about gay men (e.g., "Gay men should be allowed to adopt children") and were randomly mixed with questions concerning attitudes towards moderately related concepts (sexuality, gender-stereotypic behavior, authoritarianism, and conservatism). The final question always concerned participants' sexual orientation.

Procedure

Participants were tested individually in experimental cubicles equipped with iMacs. The presentation of the instructions, the explicit questionnaire, and the IAT was controlled by a computer program (Steffens, 1999a). The explicit attitude questionnaire was administered first. In order to minimize the influence of self-presentational factors on the responses, participants were guaranteed that their responses could not be associated with their names at any time. The questions were presented one at a time in an individual random order. Participants responded by indicating, on a 9-point scale, how much they agreed or disagreed with each statement. The final question concerned sexual orientation.

For the IAT, participants were informed that their task was to categorize words as belonging to the category displayed at the top left or right screen corner by pressing, as quickly as possible, the respective response key. There were 20 trials in each of three practice tasks (see Greenwald et al., 1998). The congruent and the incongruent task each consisted of 2 blocks of 40 trials. The first half of the participants received the *congruent*, heterosexual + positive task first. The other half of the participants first received the *incongruent*, gay + positive task. Within each task, category instances were brought into an individual random order. Category assignment to the left or right response

key was counterbalanced. The reaction–stimulus interval was 400 ms. Errors resulted in an appropriate visual feedback. Participants received feedback on errors and reaction times after each block of trials.

After the first session, participants were scheduled to return one week later (plus or minus one day) at the same time of day (plus or minus one hour). An anonymous individual code (a combination of letters in parents' names, etc.) assured that participants received the identical randomized input file again. Afterwards, participants were offered an explanation as to the purpose of the experiment.

Design

The main dependent variables were the reaction times in the IAT and the scores on the explicit attitude questionnaire. Independent variables were task congruency and measurement occasion (both within subject).

Results

In both experiments, following Greenwald et al. (1998), the first two reactions of each block of IAT trials were not analyzed and reaction times below 300 and above 3000 ms were recoded to the respective values. Reaction times associated with incorrect responses were included. For all analyses, the reaction time data were log-transformed. However, Figure 2 shows the more familiar untransformed data. The outlier treatment (see Miller, 1991; Ulrich & Miller, 1994) and the data transformation did not affect the pattern of results in the general linear model analyses of the data. Prior to structural equation modeling, the log-transformed data were carefully screened and found to correspond well to a normal distribution in both experiments reported. Additionally, bivariate screening showed that relationships between variables were linear (see Kline, 1998).

All significance tests were conducted with $\alpha < .05$, and individual p values are omitted for significant effects. The partial squared correlation, R^2_p (which captures, by definition, the relation between the variance of the predicted scores and the variance of the observed scores) is reported as an indicator of the effect size (see Cohen, 1977). Excluding the non-heterosexual participants did not change the results of the statistical tests. However, as one would expect, average attitudes towards gay men were less tolerant if analyses were performed excluding them.

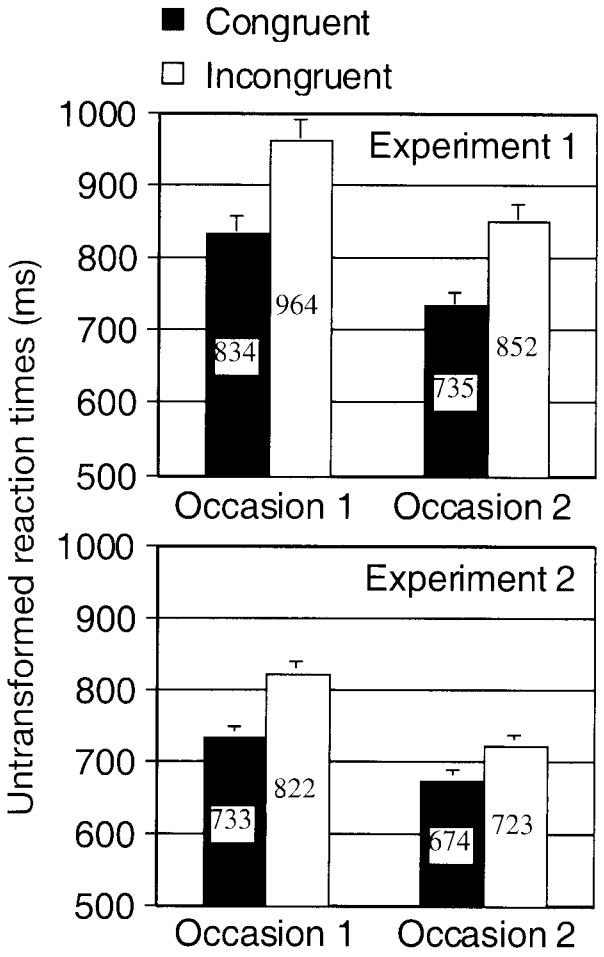

Figure 2. Mean untransformed reaction times in Experiments 1 and 2, separately for the congruent and incongruent task and for Occasion 1 and 2. Error bars reflect standard errors of means.

Implicit Attitude Measurement: Reaction Time Analyses

The average error rate was .044. The upper panel of Figure 2 shows the means of the untransformed reaction times in the congruent (heterosexual + positive) and incongruent (gay + positive) task, separately for the first (Occasion 1) and the second measurement occasion (Occasion 2). The typical IAT effect is obvious from the fact that response times were longer in the incongruent than in the congruent task. Responses were faster at Occasion 2, but the difference between the congruent and the incongruent task was not changed. A 2×2 analysis of variance (ANOVA) on the log-transformed reaction times with task congruency and measurement occasion as within-subject variables confirmed a significant effect of task congruency (the IAT effect), $F(1, 83) = 94.19$, $R^2_p = .53$, and an effect of measurement occasion, $F(1, 83) =$

Table 2. Sample Covariances (Lower Triangular Matrix Including the Diagonal) and Correlations (Upper Triangular Matrix) for the IAT Odd–Even Test Halves at Occasion 1 and 2 in Experiments 1 and 2.

	Experiment 1 (delay 1 week)			
	IAT 11	IAT 12	IAT 21	IAT 22
IAT 11	41.7231	*.7834*	*.4675*	*.4749*
IAT 12	33.2289	43.1167	*.4703*	*.3857*
IAT 21	19.7839	20.2293	42.9188	*.8161*
IAT 22	19.8068	16.3554	34.5243	41.6998
	Experiment 2 (delay 10 minutes)			
	IAT 11	IAT 12	IAT 21	IAT 22
IAT 11	37.5985	*.9160*	*.5314*	*.5418*
IAT 12	34.3146	37.3245	*.5455*	*.5264*
IAT 21	16.4305	16.8047	25.4266	*.8481*
IAT 22	15.3304	14.8399	19.7337	21.2908

136.93, $R^2_p = .62$, but no interaction between these variables, $F < 1$.

Implicit Attitude Measurement: Reliability Analyses

Internal Consistency. After computing IAT effects for the log-transformed reaction times separately for each of the 20 stimulus words (cf. Steffens & Plewe, 2001), we found a very good internal consistency of .88 and .89 (Cronbach's α) for the first and second measurement occasions, respectively.

Reliability Model. Before calculating covariances, we multiplied all differences between log-transformed response times by 100 to avoid small numbers. Test halves were created using the odd–even method. Table 2 shows the sample covariances and correlations for the four measurements; that is, the two IAT test halves at both measurement occasions (with "IAT 12" denoting the "second" test half at the first measurement occasion). The reliability model was fit to these data, testing whether the four measurements can be conceptualized as equivalent measurements of one single underlying true-score variable τ ("attitude towards gay men"). Two parameters were estimated. This classical reliability model did not fit the data, $\chi^2(8) = 73.76$, root mean square error of approximation (RMSEA) = .32 (90% confidence interval: .25–.38). The left half of the upper panel of Figure 3 shows the standardized solution of this model, with parameters estimated and variances set to 1. The right half of the upper panel depicts the same model, but with the parameters set to 1 and the variances estimated.[1] Thus, the assumption that the four IAT measurements can be conceptualized as equivalent measurements of one single underlying true-score variable must be rejected.

Stability Model. The stability model assumes that two different, but correlated true-score variables τ_1 and τ_2 generated the data at the first and second measurement occasion, respectively. Four parameters were estimated. When we fitted a variant of this model that assumed essential τ-equivalence of the four measured variables, the fit was very good, $\chi^2(6) = 6.17$, RMSEA = .02 (90% confidence interval: .00–.15). We therefore need not reject this model and the assumptions it implies. This model's most interesting parameter in the present context is the correlation between the two true-score variables τ_1 and τ_2, which may be obtained from the standardized solution. This correlation is .56 (see the left half of the middle panel of Figure 3).

Consistency Model. We also fitted the consistency model to the data. To reiterate, the consistency and stability models are data equivalent in the present situation, which implies identical model fit. However, the consistency model has the advantage of decomposing the true-score variance at each measurement occasion into components that represent the variance specific to the measurement occasion, ζ, and the variance common to the two true-score variables, ξ. This second component is a stable, transsituational component, which is obviously closer to the concept of one "trait-like" factor (a transsituationally stable attitude towards gay men in the present case) underlying the measurements than the idea of two different, but correlated factors of the stability model.

The results for the consistency model are depicted in the lower panel of Figure 3. Clearly, the proportion of the variance of both τ_1 and τ_2 accounted for by the transsituational variable ξ (19.04) is not much larger than the proportion accounted for by the situation-specific variables ζ (14.54 and 15.13), as the model on the right shows. In other words, our IAT measurements at each measurement occasion do not only contain the usual measurement error, but they also contain a sizeable component that is situation specific and not a "trait-like" attitude.

[1] The variant of the reliability model illustrated in Figure 3 not only assumes that the measured variables are the realizations of one latent variable, it also assumes that the four measured variables are essentially τ-equivalent (cf. Steyer, 1989). This is why the error components ε_{ij} were restricted to be equal. The τ-equivalence assumption is very reasonable as both test halves used the same number of items – in fact, the very same items. Nevertheless, we also fitted a variant of the reliability model in which the four measured variables were assumed to be only τ-congeneric (see Steyer, 1989). This model, in which the variances of the error components were free to vary, did not fit the data, either, $\chi^2(5) = 73.16$ (Experiment 1), and $\chi^2(5) = 150.92$ (Experiment 2).

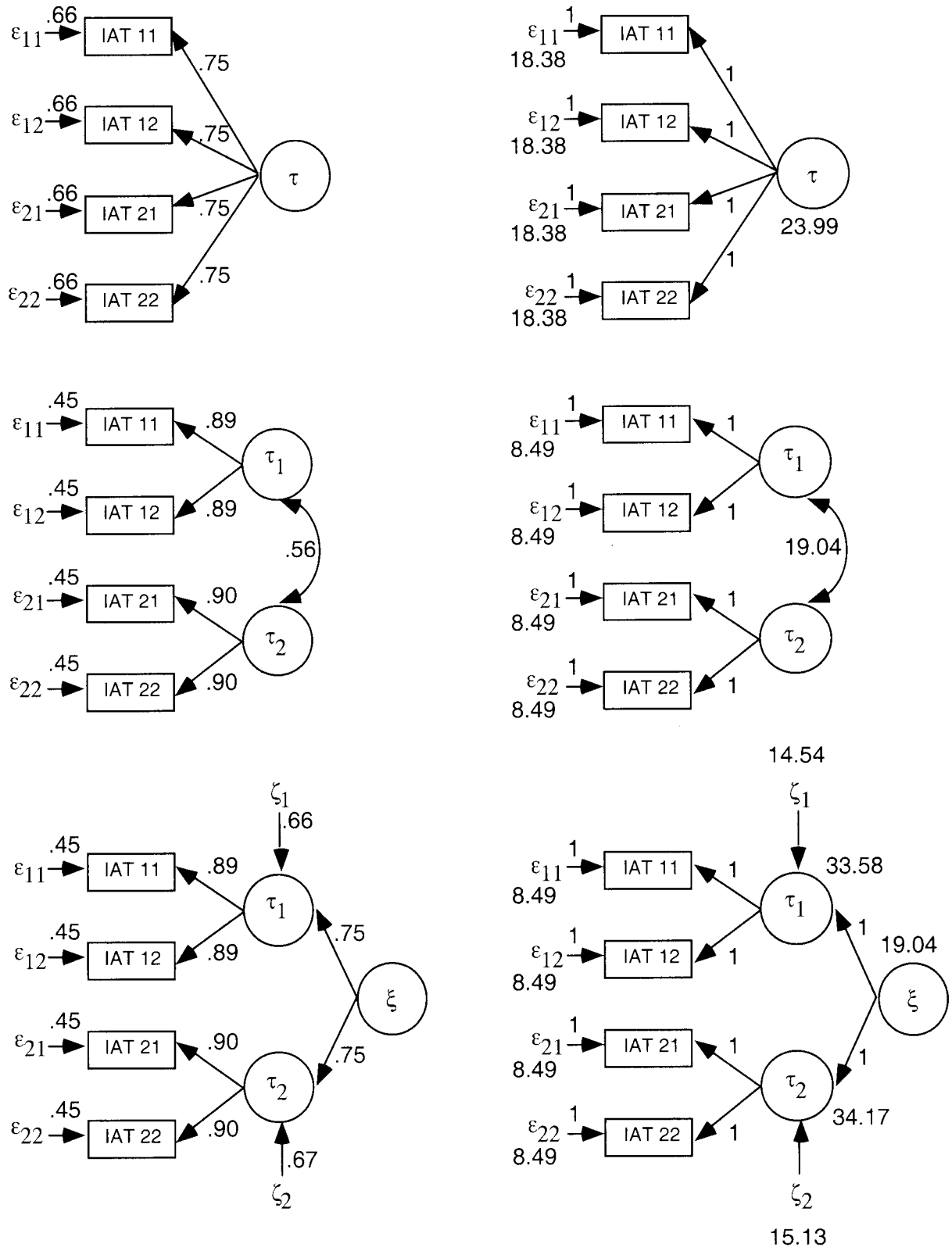

Figure 3. The reliability model (upper panel), the stability model (middle panel), and the consistency model (lower panel) applied to the data of Experiment 1. Standardized solutions with parameters estimated are depicted on the left, non-standardized solutions with variances estimated are depicted on the right.

Table 3. Partial Product-Moment Correlations Between the Implicit and Explicit Attitude Measures at Measurement Occasion 1 and 2, Controlling for Task Order.

	Experiment 1		
	IAT, Occasion 1	IAT, Occasion 2	EQ, Occasion 1
IAT, Occasion 2	**.50**		
EQ, Occasion 1	.08	**.29**	
EQ, Occasion 2	**.22**	**.44**	**.87**
	Experiment 2		
	IAT, Occasion 1	IAT, Occasion 2	ATG, Occasion 1
IAT, Occasion 2	**.52**		
ATG, Occasion 1	**.29**	**.27**	
ATG, Occasion 2	**.33**	**.29**	**.92**

Note. Significant correlations are shown in bold print.
The explicit attitude measurements were scores on a 10-item ad-hoc questionnaire (EQ, Experiment 1) and on Herek's (1994) log-transformed ATG (Experiment 2).

Explicit Attitude Measurement: Questionnaire Data

An analysis of the explicit data showed that the 10 statements measuring attitudes towards gay men had a satisfactory internal consistency (Cronbach's α = .79 and α = .77 at Occasion 1 and 2, respectively). Within a range between 10 and 90 (10 signifying the most positive attitude), the average attitude score was 42.43 (SD = 10.34) and 43.36 (SD = 9.99) at Occasion 1 and 2, respectively. In spite of that satisfactory internal consistency, model-based analyses were not possible for the explicit attitude data of both experiments reported.

Table 3 shows the correlations of the explicit and implicit attitude measures within and across measurement occasions. There were substantial correlations between the explicit and the implicit attitude measures, especially at Occasion 2. The test-retest correlation for the explicit attitude measurement was high (.87) and clearly higher than that for the implicit IAT measure (.50). In fact, a test for correlated but nonoverlapping correlations showed that these correlations were significantly different, z = −4.93 (see Raghunathan, Rosenthal, & Rubin, 1996).

Sexual orientation was indicated by each participant at each measurement occasion. The reliability of this explicit one-item questionnaire was far from perfect, Spearman's ϱ = .79.

Discussion

Participants reacted considerably faster when "heterosexual" and "positive" were assigned to the same response key than when "gay" and "positive" were assigned to one key. This suggests an implicit preference for heterosexual over gay. Further, we replicated others' findings of moderate correlations between IATs and explicit measures, and we found that the IAT effect (despite being a difference-based measure) has a very high internal consistency. Thus, the IAT properties that can be analyzed within one measurement occasion are very promising.

However, if a delay of one week between measurement occasions is introduced, the correlation between the IAT effects was rather low at the level of the measured variables. At the level of latent variables, the stability model showed that the true-score variable of the first measurement occasion did not predict much more of the variability in the true-score variable at the second measurement occasion. The consistency model revealed that slightly less than half of the variance of the true-score variables at each measurement occasion was situation-specific. This stands in sharp contrast to the explicit measure of attitudes towards gay men. At the level of the measured variables, the correlation between the questionnaire responses was much higher. This discrepancy is even more surprising if one considers that our questionnaire was constructed in a rather ad-hoc fashion and exhibited lower internal consistency than the IAT did.

One reason why explicit attitude measures appear more consistent than perhaps they really are may be carry-over effects. Either the explicit attitude is construed similarly at different points in time, or participants remember their previous replies and reply similarly in order to appear consistent. Perhaps in contrast to the explicit attitudes, a variety of as-yet-unknown situational factors influences implicit attitudes. If this were so, they would indeed more resemble the mixture of trait-like and state-like components that our models suggest than their explicit counterparts. This is plausible given the recently reported interactions of IAT effects with situational

factors (Blair et al., 2001; Dasgupta & Greenwald, 2001; Karpinski & Hilton, 2001; Kühnen et al., 2001; Lowery et al., 2001). Thus, it is perfectly reasonable that the correlation between the measurement occasions separated by a week was not higher than that observed in our Experiment 1. Many uncontrollable influences on these implicit attitudes may have been present. If this were so, higher correlations should be obtained on an immediate retest. We investigated this possibility in Experiment 2.

Experiment 2

Clearly, the most interesting finding of Experiment 1 was the small proportion of variance in implicit attitudes, measured with the IAT, that is accounted for by a transsituational "trait-like" variable given a mere week between measurement occasions and no attempt at manipulating implicit attitudes. In Experiment 2, we optimized the conditions for measuring stability in implicit attitudes (a) by reducing the interval between measurement occasions from one week to just ten minutes, and (b) by doubling the number of items to which participants responded in the IAT on the combined discrimination tasks – it could be that too few trials were administered in Experiment 1 for participants to reach a "plateau" of stable reaction time differences. A further change was that instead of using names of fictitious couples with only a learned association to the target categories, we used words that were stereotypically (and thus pre-experimentally) associated with the target categories. This was done in an attempt to make the relation between the target categories and the to-be-categorized instances more meaningful. As of now, much is unclear about the role of stimuli in the IAT (Steffens et al., 2001a). Our stereotypically associated words could only be classified in relation to the attitude targets, whereas participants may, in the course of the experiment, learn other classification rules for names (e.g., "if female, then left"), ignoring the targets. Such a strategy could lower retest correlations. In addition, rather than using an ad-hoc constructed questionnaire, a translation of the 10-item version of the Attitude Toward Gay Men Scale (ATG, Herek, 1994) was used in Experiment 2.

Method

Participants

A total of 107 students (27 male) at the University of Trier volunteered or participated for course credit without being informed beforehand about the topic of investigation. Their mean age was 23 years (SD = 3.4). According to the Kinsey scale, 21% of them were not heterosexual.

Materials

The materials were those used in Experiment 1, with the following exceptions. Instead of name pairs, items stereotypically associated with the target categories were presented, selected after pilot studies for previous research in our laboratory (Steffens, 1999b, 2002b). The average rating of the 'heterosexual' items (Casanova, Prostitution, Kirche, Kinder, Scheidung – Casanova, prostitution, church, children, divorce) was not more positive (cf. Steffens & Plewe, 2001) than that of the gay items (George Michael, Stricherszene, outen, Künstler, Tunte – George Michael, male prostitution, outing, artist, drag queen). A translation of the 10-item version of the ATG (Herek, 1994) was administered as the explicit measure of attitudes towards gay men.

Procedure

The procedure was identical to that of Experiment 1, with the following exceptions. At the first measurement occasion, the IAT and the ATG were administered. The critical part of the IAT itself was extended, resulting in 4 instead of 2 blocks of 40 trials of the congruent and incongruent IAT tasks. Participants then rated their own sexual orientation. Roughly ten minutes separated the first and second measurement occasion. The second measurement was identical to the first except that sexual orientation was not assessed again.

Participants read newspaper articles between measurements. Half of them read one article that presented a dislikable heterosexual person and another that presented a likable gay man, the other half read two articles unrelated to the topic of investigation. Which articles participants read did not affect the implicit and explicit attitude measures, their psychometric properties, correlations, or covariances in any detectable way. Therefore, the variable is omitted in the present article.

Design

The main dependent variables were the reaction times in the IAT and the scores on the explicit attitude questionnaire, the ATG. The independent variables were task congruency and measurement occasion (both within subject).

Results

Implicit Attitude Measurement: Reaction Time Analyses

The average error rate was .048. The lower panel of Figure 2 shows the untransformed reaction times in congruent and incongruent tasks, separately for the first and the second measurement occasions. The IAT effect (the difference between congruent and incongruent trials) was larger at Occasion 1 than at Occasion 2.

A 2 × 2 ANOVA on the log-transformed reaction times with task congruency and measurement occasion as within-subject variables confirmed that there was a significant IAT effect, $F(1, 106) = 59.67$, $R^2_p = .36$, an effect of measurement occasion, $F(1, 106) = 140.56$, $R^2_p = .57$, and an interaction, $F(1, 106) = 8.46$, $R^2_p = .07$. Tests of simple main effects showed an IAT effect at Occasion 1, $F(1, 106) = 53.56$, $R^2_p = .34$, and at Occasion 2, $F(1, 106) = 39.28$, $R^2_p = .27$.

Implicit Attitude Measurement: Reliability Analyses

Internal Consistency. Cronbach's α for the IAT scale was .93 and .90 for the first and second measurement occasions, respectively. Thus, the internal consistency of the IAT was exceptionally good and even somewhat higher than in Experiment 1 – which is to be expected given the doubled number of reactions.

Reliability Model. The lower half of Table 2 shows the sample covariances and correlations for the IAT test halves. As in Experiment 1, the reliability model (see the upper panel of Figure 4) did not fit the data, $\chi^2(8) = 170.15$, RMSEA = .44 (90% confidence interval: .38–.50).

Stability Model. In contrast to the reliability model, the stability model, assuming essentially τ-equivalent measured variables, fit the data very well, $\chi^2(6) = 5.31$, RMSEA < .01 (90% confidence interval: .00–.12) (see the middle panel of Figure 4). The correlation between the two true-score variables τ_1 and τ_1 ("attitude towards gay men at Occasion 1 and 2") was .61 (see the model on the left). This is a negligible improvement over the same correlation reported for Experiment 1 (.56), considering that the inter-test interval was reduced from 1 week to just 10 min.

Consistency Model. The results for the consistency model are depicted in the lower panel of Figure 4. The results for the first measurement occasion are similar to those of Experiment 1 in that a large proportion of the variance of the true-score variable τ_1 is accounted for by situation-specific effects (18.34; see the model on the right). In fact, the proportion of the variance of τ_1 that is accounted for by the transsituational factor ξ (15.85) is even somewhat smaller than the proportion accounted for by the situation-specific factor ζ_1. In contrast, a large amount of the variance of τ_2 (which is smaller than that of τ_1), almost 80%, is accounted for by the transsituational factor.

Explicit Attitude Measurement: ATG Data

Cronbach's α of the 10-item ATG was .83 and .85 for the first and second measurement occasion, respectively, indicating that the internal consistency of this measurement instrument was acceptable. However, an analysis of the summary score computed across the 10-item ATG showed that this score was not distributed normally (skewness > 1.21). This was due to a noticeable floor effect in that our participants preferred the egalitarian end of the rating scale (see also Steffens, 2002b). We log-transformed the scores to obtain a better approximation to a normal distribution. Means of the untransformed/log-transformed data were 24.79 (*SD* = 12.52)/3.10 (*SD* = .47) and 24.34 (*SD* = 13.12)/3.06 (*SD* = .51) at Occasion 1 and 2, respectively.

All correlations between explicit and implicit attitude measures were of almost medium size according to Cohen's (1977) standards (see Table 3). The test-retest correlation for the explicit attitude measurement was again very high (.92) and significantly higher than the test-retest correlation for the implicit measure (.52), $z = -7.30$.

Discussion

The IAT and the ATG differed drastically with respect to their stability. The plain test-retest correlations are already quite informative. In spite of the fact that scores on the 10-item ATG were so skewed that they had to be log-transformed in order to show an acceptable approximation to a normal distribution, log-transformed scores from the first and second measurement occasion correlated highly. This may not be awfully surprising considering that only a few minutes separated the occasions. However, the 152-item IAT effects from the first and second measurement occasion correlated rather poorly. Even at the level of latent variables, the stability model showed that the correlation between the true-score variables of the first and second measurement occasion was rather low. Note that this is so even though the internal consistency of the IAT was very good. Again, there were medium correlations between the IAT and explicit measures.

In contrast to Experiment 1, in Experiment 2 we found a smaller IAT effect at the second as opposed to the first measurement occasion. Smaller IAT effects in a second IAT administered immediately after

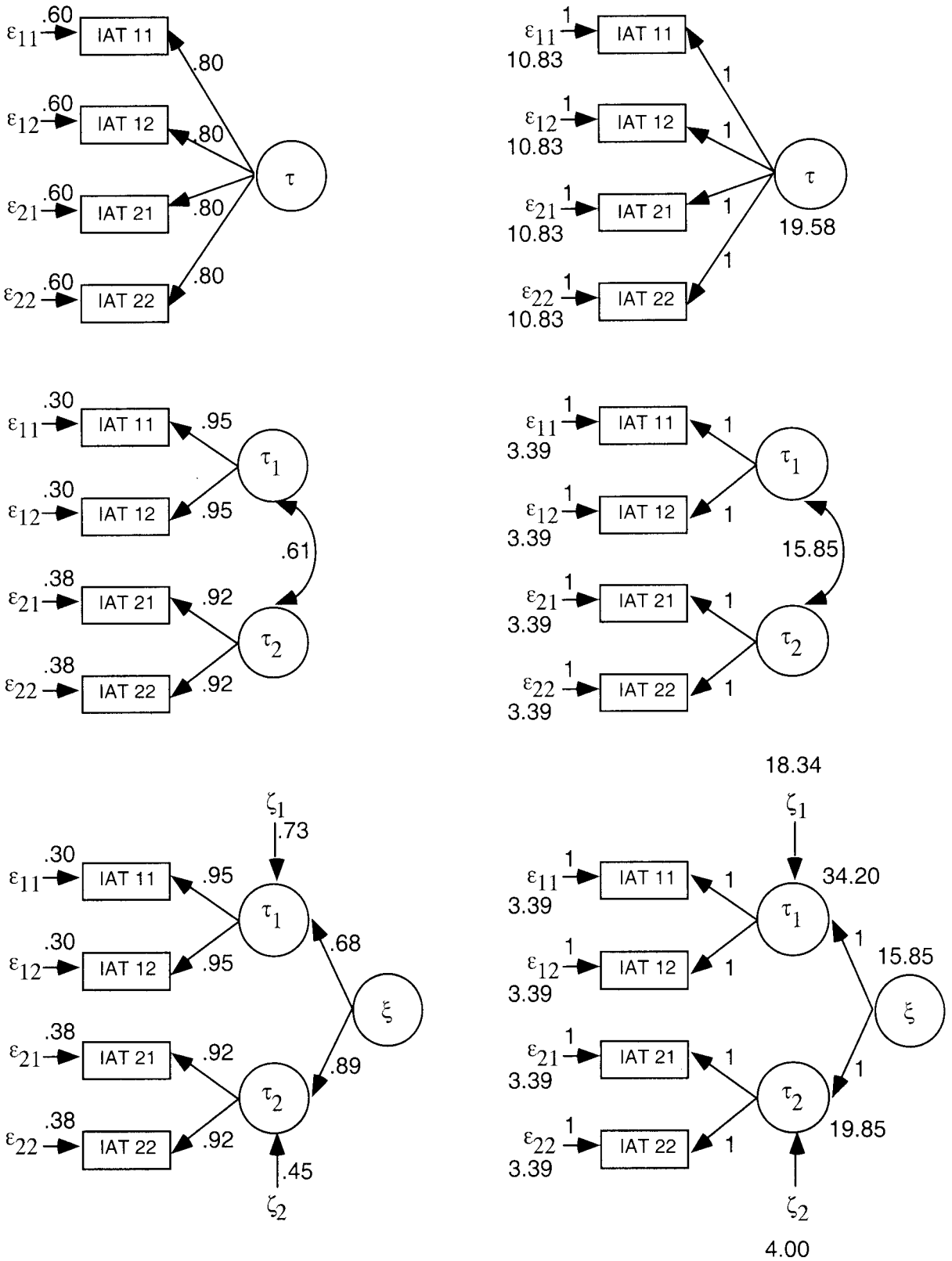

Figure 4. The reliability model (upper panel), the stability model (middle panel), and the consistency model (lower panel) applied to the data of Experiment 2. Standardized solutions with parameters estimated are depicted on the left, nonstandardized solutions with variances estimated are depicted on the right.

a first, related one have been observed in other, as yet unpublished experiments in our laboratory. The shortened delay between tests thus seems responsible for this finding. In line with this suspicion, in Experiment 1, too, we found a larger IAT effect for the first as opposed to the second blocks administered at each measurement occasion, with a similar-size interaction as for IATs in Experiment 2 ($F(1, 81) = 4.58$, $R^2_p = .05$).

In sum, the results of Experiment 2 replicate those of Experiment 1. In contrast to that experiment, there is little reason to believe that meaningful changes in implicit attitudes occurred between the measurements, given the small delay in Experiment 2. Both experiments together nourish the suspicion that what the IAT measures is not confined to a trait-like, transsituationally stable, individual implicit attitude.

General Discussion

"The mediating role of one's attitudes on one's behavior moved from being described in terms of a conscious and intentional retrieval of one's attitude from memory, to a demonstration of automatic attitude activation and influence," summarized Bargh (1997, p. 5). Consequently, "efficient assessment of individual differences in implicit social cognition" seemed to be "perhaps the most significant remaining challenge" in social cognition research a few years ago (Greenwald & Banaji, 1995, p. 20). This goal can only be obtained if a stable, person-related factor in implicit attitudes can be identified and measured reliably. Our experiments focused on the stability of implicit attitudes towards gay men measured with an IAT. IAT scores were compared to (1) the same scores on a delayed retest, (2) the same scores on an immediate retest, and (3) scores on explicit attitude tests. The within-occasion internal consistency and the split-half correlation of the IAT were remarkable, especially given the fact that IAT scores are difference scores. Indeed, IAT scores were clearly superior to those on the explicit tests we used. We also found medium correlations with attitudes as measured in those explicit tests. These results join the emerging canon of findings indicating the IATs' validity for attitude measurement – when data are evaluated at a group level. However, the main finding of our experiments is that the test-retest correlation of an IAT assessing implicit attitudes, as obtained directly or estimated in a structural equation model, was rather low ($.50 < \hat{r} < .62$), even when the IAT was replicated immediately.

One might suspect that the low test-retest correlations are due to the fact that our sample was relatively homogenous with regard to the attitudes they held – at least if we can trust the explicit tests to reveal the "true" attitudes (which, of course, we do not). If efforts were made to select a sample ranging from extremely positive to extremely negative attitudes towards gay men, larger test-retest correlations may be found. In addition, randomization and counterbalancing may have lowered correlations (see Banse et al., 2001). Although this is possible in principle, we do not believe it to play a major role because we did not simply find low correlations. We found that the IAT's test-retest correlation was much lower than the split-half correlation, and also much lower than the test-retest correlation for the explicit test. In other words, testing was highly reliable at any given point in time, or with an explicit measure, despite possibly restricted variance in the sample. But what was measured on the implicit task seems to differ from one testing to the next, even if measurement occasions were only a few minutes apart.

We changed several aspects simultaneously between Experiments 1 and 2. We believe that the cross-experimental consistency of the findings points to their generality: Rather low test-retest correlations were found independently of the length of the IAT, of the concept associates used, and of the delay between measurement occasions. Thus, the transsituational factor that we found seems very stable. However, it may also be possible that several effects cancelled each other out, and that a more substantial transsituational factor would be found with some other combination of these factors. As a sidenote, we tried several ways of "cleaning up" the data in order to increase the test-retest correlation (e.g., exclude the first block of trials in each task), but in vain.

There may be another feature of the IAT that could reduce its reliability. Practice resulted in faster reactions on the second IAT than on the first. Participants may benefit from practice to different degrees, and some may benefit more in incongruent tasks than in congruent ones. If this were so, correlations between IAT effects in adjacent blocks of trials should be higher than correlations between IAT effects in nonadjacent blocks of trials. Consistent with this assumption, the correlation between the IAT effect in the second block of the first IAT and the adjacent IAT effect in the first block of the second IAT was .53 in Experiment 1. Instead, the correlation between the IAT effect in the first block of the first IAT and the IAT effect in the second block of the second IAT was .18, which is significantly lower, $z = 3.01$. Thus, there seems to be gradual rather than abrupt change in IAT effects over practice. Given this finding, which measurement is to be trusted? That based on early, nonadapted behavior or that based on the "plateau" that a particular participant reaches? A definite answer to this question is beyond the scope of the present article, but this answer is important if the IAT is ever to be applied outside the laboratory. For the data of Experiment 2, we calculated IAT ef-

fects based on subsets of the eight blocks of IAT trials, and we compared their correlations (controlling for task order) with explicit attitudes, participant sex, and participant sexual orientation. Numerically, the highest correlations were found for the IAT effect summed over Blocks 2 to 7 (e.g., the obtained correlations with the ATG were .33 and .36 at Occasion 1 and 2, respectively). While the increase in correlations was moderate, we conclude, in line with the difference in internal consistencies between Experiments 1 and 2, that more reliable data are obtained in longer as opposed to shorter IATs. These analyses suggest that there is gradual change in the size of the IAT effect over the course of an IAT (or several IATs). The IAT effects over the first dozens of reactions (which are routinely excluded from analyses) seem to be the least related with (explicit) attitudes. The subsequent reactions are in part explained by explicit attitudes and behavior, but they still contain a method factor related to learning that best cancels out if many data points are analyzed.

For IATs, most other authors have reported test-retest correlations in the same order of magnitude as those found here (see Table 1). When compared to established tests of personality traits (e.g., Steyer et al., 1989), the transsituational component of IATs looks quite small. Cunningham et al. (2001) carried out a series of experiments that pursued similar aims as the research reported in the present article. The most important difference between their approach and ours is that we modeled situational factors in addition to the transsituational one by assessing test-retest correlations as compared to split-half correlations. Cunningham et al., observing standardized weights of .38 to .73 on the IAT-related paths, arrived at the conclusion that "after correction for measurement error, implicit attitude measures proved consistent across time and across measures" (p. 169).

However, it depends on the purpose of measurement how high a standard one must impose on a measurement's reliability (see Brown, 1983; Pedhazur & Schmelkin, 1991). For individual diagnosis, very high standards are typically required. Whereas some authors consider a test-retest-correlation above .60 "marginally acceptable" (Gliner & Morgan, 2000, p. 313), others go as far as suggesting levels of .85 or even .90 (Brown, 1983; Kelley, 1927) as acceptable. If we posit that the correlation between the true-score variables should be at least .80, and if we fit a variant of the stability model to the data with the correlation between τ_1 and τ_2 set to .80, we arrive at a significant reduction in model fit, as compared to the model without that restriction, $\chi^2(1) = 11.71$ and $\chi^2(1) = 12.83$ in Experiment 1 and 2, respectively. That is, our data are not compatible with the assumption that the correlation between the true-score variables is .80. However, one might argue that such high standards apply only to fields where established measures are used for diagnostic purposes in applied contexts. Implicit attitude measurement certainly is not such a field. Thus, Greenwald and Farnham (2000) propose that an IAT test-retest correlation of .55 is satisfactory *for research purposes*, and we agree: Reliabilities in that order of magnitude may be considered sufficient for assessing differences between groups.[2]

An illustration of our results with a focus on individual scores can be found in Table 4. We *z*-transformed the IAT effects obtained in Experiments 1 and 2 at each measurement occasion and recoded them onto a 5-point scale ranging from 'very negative' to 'very positive' so that roughly the same number of participants (between 15% and 25%) fell into each of those five intervals. Take Experiment 2: Less than one third of our 107 participants would have received the same diagnosis about their "implicit homophobia" as a few minutes before. Only half of the 20 participants who would certainly have been singled out for "homophobia training" on the basis of their Occasion 1-scores seemed that homophobic at Occasion 2. However, 12 other participants "became" very homophobic at Occasion 2. For comparison, the log-ATG scores of two thirds of our participants, similarly transposed onto a 5-point scale, were identical on Occasion 1 and 2.

We are of one voice with other research groups in claiming that, beyond any doubt, a transsituational attitude-related factor is measured by an IAT. However, this stable factor is not dominant enough to warrant interpretations of IAT effects at the level of the individual. Our data make a very strong point against using individual differences in IAT scores for diagnosis or intervention, and we think the data presented in Cunningham et al. (2001) should be interpreted accordingly. Either our IAT is not a reliable enough measure-

[2] Our model-based analyses suggest a lot of change in implicit attitudes from one measurement occasion to the next. One may always object against model-based approaches that the conclusions depend on the specific model used and the assumptions implied. Therefore, as an alternative, a coefficient of reference reliability, r_{rf}, was computed for the IAT effect. This coefficient relates variability that is due to change on the one hand to variability due to change and error combined on the other (Schweizer, 1988). Change is estimated by means of the test-retest correlation coefficient, error is estimated by means of the correlation coefficient of the combined odd–even halves of both tests. A resulting coefficient of $r_{rf} = .93$ for Experiments 1 and 2 indicated a high degree of reference reliability, that is, much change and little error in IAT effects from the first to the second measurement occasion. This was found despite the fact that measurement occasions were in such temporal proximity in Experiment 2. These supplementary analyses corroborate our model-based findings of a substantial situational factor.

Table 4. "Attitudinal" Transitions From the First Measurement Occasion to the Second for a 5-Point Implicit Attitude Scale Ranging From Negative to Positive, for Experiments 1 and 2.

Experiment 1 *Implicit attitude, Occasion 1:*	*Implicit attitude, Occasion 2:* Negative	Moderately negative	Neutral	Moderately positive	Positive
Negative	6	2	6	3	0
Moderately negative	4	4	4	1	2
Neutral	3	6	6	2	2
Moderately positive	1	3	1	5	6
Positive	1	2	4	4	*6*
Experiment 2 *Implicit attitude, Occasion 1:*	*Implicit attitude, Occasion 2:* Negative	Moderately negative	Neutral	Moderately positive	Positive
Negative	10	6	3	0	1
Moderately negative	7	8	4	1	3
Neutral	3	5	5	9	5
Moderately positive	1	4	6	4	3
Positive	1	1	1	8	*8*

Note. The number of participants with "stable attitudes" is presented in italics.

ment instrument for individual diagnosis, or there is more to implicit attitudes than a stable, trait-like component (see Kühnen et al., 2001), or both.

References

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

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *The automaticity of everyday life: Advances in social cognition, Vol. 10* (pp. 1–61). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81,* 828–841.

Bohrnstedt, G. (1993). Classical measurement theory: Its utility and limitations for attitude research. In D. Krebs & P. Schmidt (Eds.), *New directions in attitude measurement* (pp. 169–186). Berlin: de Gruyter.

Bollen, K. A., & Long, J. S. (Eds.) (1993). *Testing structural equation models.* Thousand Oaks, CA: Sage Publications, Inc.

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.

Brown, F. G. (1983). *Principles of educational and psychological testing.* New York: Holt, Rinehart and Winston.

Buchner, A., & Brandt, M. (in press). Further evidence for systematic reliability differences between explicit and implicit memory tests. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology.*

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40,* 227–259.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800–814.

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.

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

Dovidio, J. F., Kawakami, K., & Beach, K. R. (in press). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology (Vol. 4: Intergroup relations).* Oxford: Blackwell.

Fernald, J. L. (1995). Interpersonal heterosexism. In B. Lott & D. Maluso (Eds.), *The social psychology of interpersonal discrimination* (pp. 80–117). New York: The Guildford Press.

Gliner, J. A., & Morgan, G. A. (2000). *Research methods in applied settings: An integrated approach to design and analysis.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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 für Experimentelle Psychologie, 48,* 85–93.

Hager, W., & Hasselhorn, M. (Eds.) (1994). *Handbuch deutschsprachiger Wortnormen* [Handbook of German word norms]. Göttingen, Germany: Hogrefe.

Herek, G. M. (1994). Assessing heterosexuals' attitudes toward lesbians and gay men: A review of empirical research with the ATLG scale. In B. Greene & G. M. Herek (Eds.), *Lesbian and gay psychology: Theory, research, and clinical applications. Psychological perspectives on lesbian and gay issues, Vol. 1* (pp. 206–228). Thousand Oaks, CA: Sage.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

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

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book Co.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

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

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81,* 842–855.

Meier, B., & Perrig, W. J. (2000). Low reliability of perceptual priming: Its impact on experimental and individual difference findings. *The Quarterly Journal of Experimental Psychology, 53A,* 211–233.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology, 4A,* 907–912.

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

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1,* 178–183.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57,* 743–762.

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.

Schweizer, K. (1988). Reference-reliability as a concept of reliability of change in times series data. *Educational and Psychological Measurement, 48,* 603–613.

Steffens, M. C. (1999a). Mac-IAT [Computer program]. ftp://cogpsy.uni-trier.de/pub/Mac-IAT/ Trier, Germany: University Trier.

Steffens, M. C. (1999b). "Wie homophob sind Sie, auf einer Skala von 1 bis 7?" – Die Erfassung der Einstellung zu Schwulen und Lesben ["What's your homophobia score, on a scale from 1 to 7?" – Assessing the attitudes towards gay men and lesbians]. In W. Köhne (Ed.), *Lesben und Schwule in der Arbeitswelt* (pp. 102–132). Berlin: Deutsche AIDS-Hilfe.

Steffens, M. C. (2002a). Faking implicit and explicit personality tests. *Manuscript submitted for publication.*

Steffens, M. C. (2002b). *Implicit and explicit attitudes towards lesbians and gay men.* Manuscript submitted for publication.

Steffens, M. C., Banaji, M. R., Jelenec, P., Wender, K. F., Anheuser, J., Goergens, N. K., Hülsebusch, T., Lichau, J., & Still, Y. (2001). *A two-factor account of the IAT effect.* Invited presentation, Psychology Department, University of Washington, Seattle, WA, USA.

Steffens, M. C., Günster, A., & Mehl, B. (2001). *Feminization of management and backlash against agentic women: A replication in Germany?* Invited presentation, Psychology Department, Rutgers University, Piscataway, NJ, USA.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie, 48,* 123–134.

Steffens, M. C., & Wagner, C. (2002). *Attitudes towards lesbians, gay men, bisexual women, and bisexual men in Germany.* Submitted for publication.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3, 25–60.*

Steyer, R., Majcen, A. M., Schwenkmezger, P., & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research, 1, 281–299.*

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123,* 34–80.

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

Received April 3, 2002
Revision received May 14, 2002
Accepted May 14, 2002

Melanie Steffens

FB I – Psychologie
University Trier
D-54286 Trier
Germany
Tel.: +49 651 201 2017
Fax: +49 651 201 2955
E-mail: steffens@uni-trier.de