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-

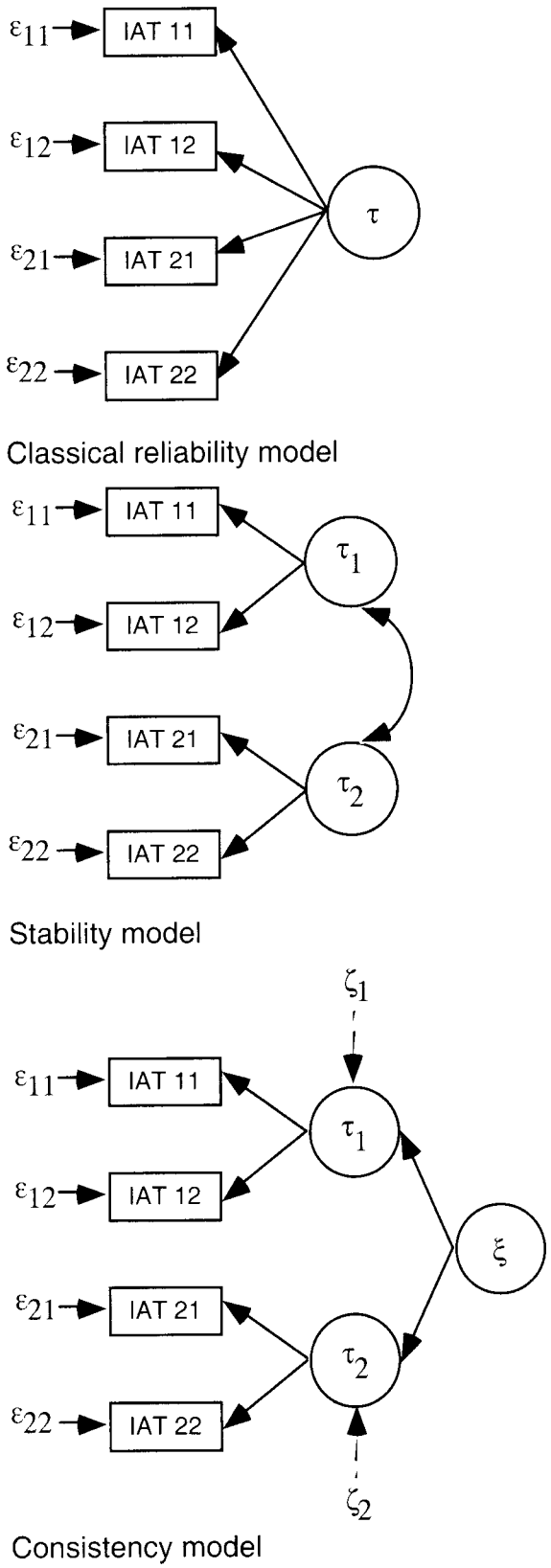

Figure 1. Three structural equation models which were fitted to the data of Experiments 1 and 2. Measured variables are displayed in rectangles, latent constructs, in circles (see text for details).

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.

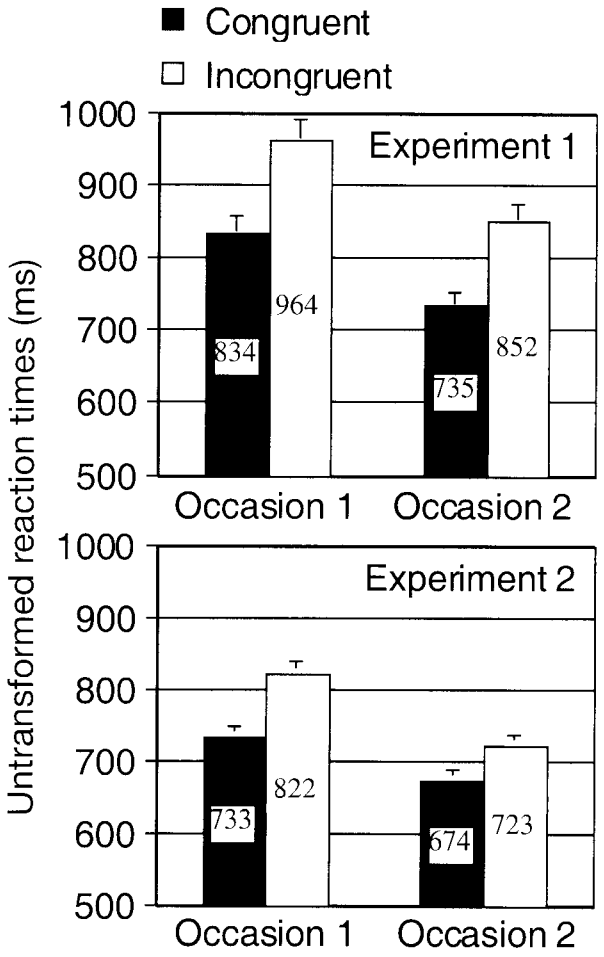

Figure 2. Mean untransformed reaction times in Experiments 1 and 2, separately for the congruent and incongruent task and for Occasion 1 and 2. Error bars reflect standard errors of means.

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).

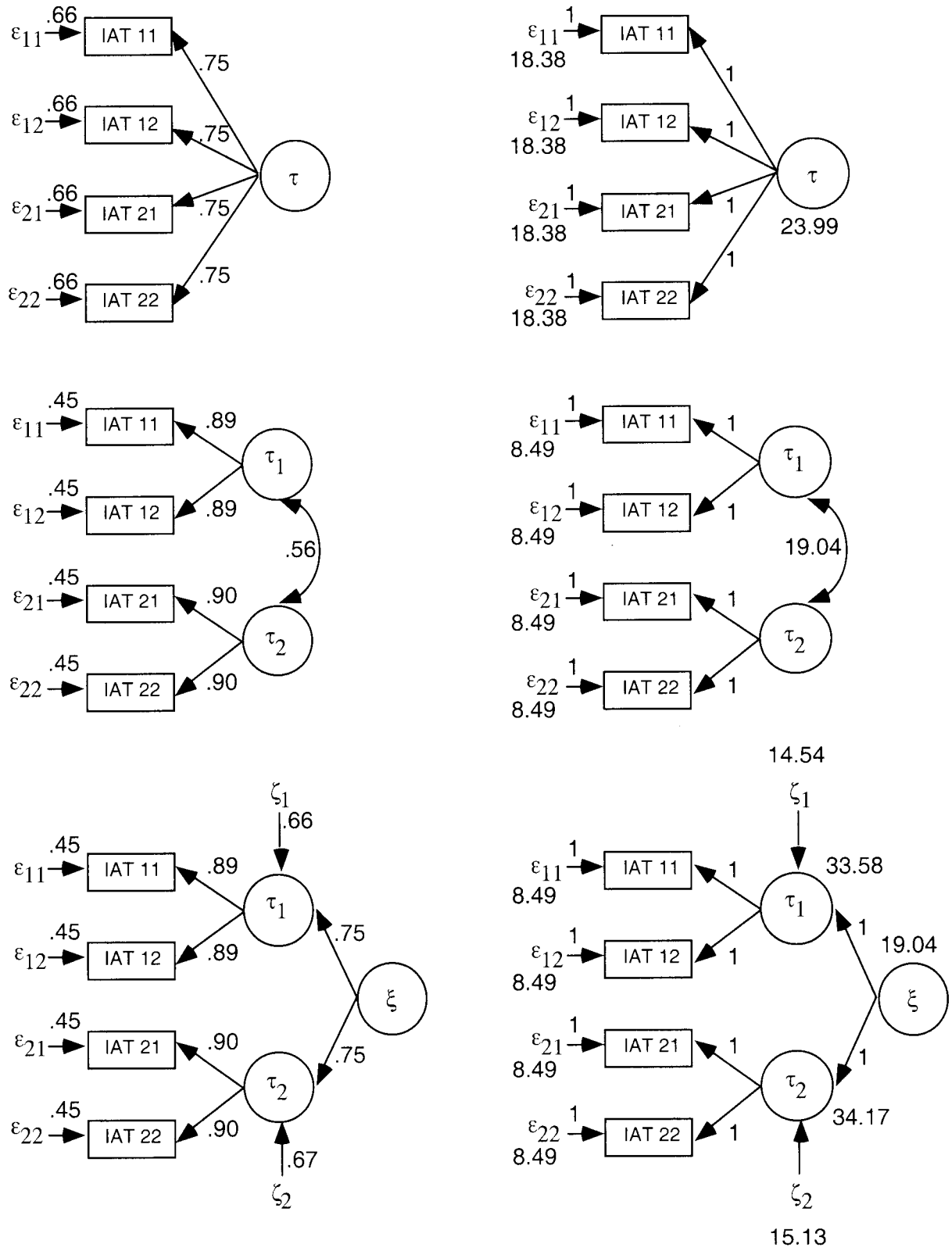

Figure 3. The reliability model (upper panel), the stability model (middle panel), and the consistency model (lower panel) applied to the data of Experiment 1. Standardized solutions with parameters estimated are depicted on the left, non-standardized solutions with variances estimated are depicted on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

