RUNNING HEAD: RELIABILITY OF IAT

Reliability Generalization of the Implicit Association Test

Sunthud Pornprasertmanit

Illinois State University

Author Notes

This article was written in April, 2009, as the term paper in Test Theory, Department of Psychology, Illinois State University.

Correspondence to Sunthud Pornprasertmanit Email: psunthud@ku.edu

Reliability Generalization of the Implicit Association Test

Wilson, Lindsey, and Schooler (2000) proposed that attitude process was classified in two parts: explicit and implicit. Explicit attitude was one which people were not able to introspect and report explicitly. Therefore, self-report attitude scale measured explicit attitude. However, implicit attitude was one which people were not aware of their preferences toward attitude targets. Implicit attitude measures were instruments which measured preferences toward target without test examinee’s introspection. Interestingly, implicit and explicit attitudes may not predict the same behaviors, which Wilson et al. (2000) represented as the dual processes of attitudes. For example, Caucasian people may have neutral attitude toward African-American people explicitly, but reveal the negative attitude when measured it implicitly (Greenwald, McGhee, & Schwartz, 1998).

Because of this interesting result and availability of implicit attitude measurement, implicit attitude has been popular since past decade. The research studies revealed the complexity of implicit attitudes. The implicit attitude studies did not provide convergent results. For example, correlations between implicit and explicit attitude measurement ranged from mildly negative to highly positive (Greenwald, Poehlman, Uhlmann, & Banaji, in press). Although the different research results may be resulted from complexity of this knowledge, the differences may come from research artifacts. One of the most important artifacts is reliability of implicit measurements, especially internal consistency. Implicit attitude measures generally have low reliability estimates compare to explicit attitude measures (Nosek, Greenwald, & Banaji, 2007).

The different reliability values of implicit measures cause different research results. Lower reliability makes standard error higher and then power of statistical testing lower. In addition, low reliability attenuates the magnitude of effect, such as correlation coefficient.

Therefore, the incongruence results of implicit attitude studies may come from the difference in reliability coefficients.

Because of the importance of reliability of implicit attitude measures, the purpose of this study is to do a quantitative review about reliability of implicit measures. I focus on implicit association test (IAT) only. The reason is that IAT is one of the most popular implicit attitude measures and it provided the highest internal consistency among implicit measures (Nosek, et al., 2007). The synthesis of IAT reliability provides the more conclusive knowledge about its quality. Moreover, this synthesis will provide some recommendations about factors which create a more reliable IAT.

Before discussing about procedure of this study, procedure of IAT and some variation of IAT will be explained. After that, I will summarize methods used for estimating IAT reliability. Next, I will hypothesize factors which possibly affect reliability of IAT based on analysis of IAT procedure and previous research studies.

Procedure of IAT

The basic idea of IAT is that time to process relating concepts together, such as positive words and flowers, is not large compared to time to process opposing concepts, such as positive words and insects. In other words, when people shift from one concept to another opposing concept, switching time is large. IAT was designed to capture this switching time between two concepts. IAT compare time people process between two attitude concepts, such as Whites and Blacks, and two opposite attributes, such as positive or negative words. If the switching time between Whites-positives and Blacks-negatives were lower than Whites-negatives and Blacks-positives, people have Whites and positive concepts related together. People may not realize this association. Some people were surprised when they know IAT score (Lane, Banaji, Nosek, & Greenwald, 2007).

To explain in detail, the IAT's procedure has seven steps. First, participants are requested to press keyboard either by left or right hand side for responding to stimuli which can be classified in two superordinate categories of attitude targets. For example, participants pressed the left and right key when seeing the Whites and Blacks faces, respectively. Second, participants are requested to press left or right key to two different groups of attribute stimuli, such as responding to positive and negative words by left and right hand sides, respectively. In the third and fourth steps, the first and second tasks are combined. For example, participants respond to either positive words or White faces by left hand and either negative words or Black faces by right hand. The third step is practice trail while the fourth step is test trial which the reaction times were used for computing IAT scores. Fifth, the first task is reversed. For example, participants respond to White faces by right hand and Black faces by left hand. In the sixth and seventh step, the second and fourth steps are combined. The pattern of stimuli matching is opposite to the third and fourth step. For example, participants respond to either positive words or Black faces by left hand and either negative words or White faces by right hand. The sixth is practice trial and the seventh step is the test trial. IAT's procedures are summarized in Table 1.

IAT will focus on the reaction time in step 4 and 7. Basically, when the stimuli were associated in the same direction, the reaction will be shorter. Therefore, if the step 4 reaction times were shorter than step 7, participants have positive attitude toward the category which coupled with positive words in step 4, and vice versa. There are many methods for computing IAT scores from these reaction times, such as differences between mean of step 4 and 7 score, difference in log-transformed score, and the D scoring algorithm (Greenwald, Nosek, & Banaji, 2003).

Another variation of IAT is single target IAT (Karpinski & Steinman, 2006). Instead of using two target categories, only one category was used to associate with attribute categories. Single target IAT was developed to solve the problem of IAT score interpretation. The IAT score should be interpreted as difference between attitudes of two categories, such as attitude difference between Whites and Blacks. It cannot be interpreted as attitude toward a single category. The procedure of single target IAT was dropped comparing category and focus on only desired target, such as measuring attitude toward Blacks by dropping Whites stimuli.

The single target IAT procedure is paralleled to IAT. Step 1 and 5 were dropped because of no opposite target category. The procedure of single target IAT starts at step 2, which is the same as IAT. In step 3 and 4, participants are requested to respond by left hand side when a stimulus is either target or one attribute, such as a positive word, and by right hand side when a stimulus is another attribute, such as a negative word. Next, in step 6 and 7, the pattern is reversed. Participants change to respond to attitude target by right hand side, while they still respond to two attribute categories in the same pattern. For example, participants respond to a positive word by left hand side and either a negative word or a target stimulus by right hand side. This study will include both IAT and single target IAT. There are other variations of IAT tasks, such as Go vs. No-go task, which I will summarize and report their reliability, but will not use these coefficients in data analysis.

Internal Consistency of IAT

Because score of IAT is based on the difference between reaction times of step 4 and step 7, coefficient alpha cannot be computed directly. There are many approaches dealing with the difference score: reliability of difference scores, split-half reliability, coefficient alpha from item parcels, and coefficient alpha of difference scores from same stimulus. First, reliability of difference scores formula (Crocker & Algina, 1986) was used, such as Cunningham, Preacher, and Banaji (2001). To estimate the reliability, the reaction times from stimuli of each step were used to find averages, standard deviations, and coefficient alphas of scores from both steps, as well as correlation of both scores. After that, calculate for reliability of difference scores by

$$r_{XX} = \frac{r_{44}s_4^2 + r_{77}s_7^2 - 2r_{47}s_4s_7}{s_4^2 + s_7^2 - 2r_{47}s_4s_7} \qquad (1)$$

r_{XX} is the reliability of IAT’s score. r_{44} and r_{77} is the reliability of score from step 4 and 7, respectively. r_{47} is the correlation between summate scores from step 4 and 7. s_4 and s_7 are the standard deviation of score from step 4 and 7. This method was calculated reliability of difference of the sum scores regardless of the fact that the reaction times toward each word or stimulus were correlated from each other.

The second method is split-half reliability. To estimate split-half reliability for IAT, the test trials are separated into two halves and calculated IAT scores for each half by the methods described above. The method of separation should produce two paralleled halves. Next, split-half reliability is estimated by various formulas. The most popular method is to find correlation of two halves and the correlation is corrected for test length by the Spearman-Brown Prophecy Formula. The result is greater than the original correlation between two halves. Another formula is Rulon’s formula. This formula is based on variability of difference scores between two halves. The Spearman-Brown Correction and Rulon’s formula will show similar reliability coefficient if two halves have the same variance (Crocker & Algina, 1986). However, if two halves variances are different, the Spearman-Brown formula creates a greater estimate.

Next, reliability is estimated by coefficient alpha from item parcels. Initially, reaction times from trial test are divided into groups, such as two (halves), three, or four. Then, the IAT scores of each group are calculated. The item parcels scores were used as item scores and calculated by coefficient alpha. The coefficient alpha is a lower bound of true reliability of a test, if the items are not strictly paralleled. Also, the coefficient alpha was proved as the average of all possible split-half reliability by Rulon's method (Crocker & Algina, 1986). Therefore, this method will generate a lower reliability estimate compared to split-half reliability when paralleled halves were created.

Sometimes, when researchers used factor analysis from item parcels, reliability can be estimated from factor loadings and error variances (Lattin, Carroll, & Green, 2003) as

$$r_{XX} = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \varepsilon_i^2} \qquad (2)$$

λ_i is the factor loading of item parcel i. ε_i^2 is the error variance of item parcel i. This reliability will create the larger estimate than coefficient alpha because it does not require the assumption of true score equivalent or equal factor loadings like coefficient alpha.

However, the problem of dividing test trials to halves or item parcels is subjectivity of a method in dividing test trials, such as odd-even or consecutive subtests. To overcome this problem, the IAT scores of the same stimulus can be created if both step 4 and 7 use the same set of stimuli. For example, if there are 40 test trials for each step and all attribute and target categories have five stimuli, the twenty item scores can be created. Next, the coefficient alpha can be applied for these item scores directly. Although this method will provide a lower reliability estimate than split-half reliability, it solves the subjectivity problem. Also, this formula will generate the better estimate than the first formula. Provided derivation in Appendix A, this method can be calculated by

$$r_{XX} = \frac{r_{44}s_4^2+r_{77}s_7^2-2r_{47}s_4s_7+\sum_{i=1}^{k} r_{(47)_i}s_{4_i}s_{7_i}}{s_4^2+s_7^2-2r_{47}s_4s_7+\sum_{i=1}^{k} r_{(47)_i}s_{4_i}s_{7_i}} \qquad (3)$$

$r_{(47)_i}$ is the correlation of mean reaction time of stimulus *i* between step 4 and 7. s_{3_i} and s_{5_i} are the standard deviation of mean reaction time of stimuli *i* in step 4 and 7, respectively. This formula is similar to the reliability of difference scores formula, except both numerator and denominator are added by function of correlation between reaction times of each stimulus. If the correlations between reaction times from step 3 and 5 of each item are positive, the formula 3 will provide greater coefficient than the formula 1. While correlation from the same stimulus was usually positive, the formula 1 will underestimate reliability coefficient.

Unfortunately, the last method cannot be used if the stimuli in step 4 and 7 are not the same. This problem occur frequently in single target IAT because researchers try to equivalent the number of trials pressed by left and right hand sides. Therefore, the numbers of positive and negative stimuli in each step were not equal, which make some stimuli appear on one step, but not other step. The easiest way is dropping the unmatched stimuli and computed coefficient alpha and do not use the unmatched stimuli for computation of IAT score.

Finally, the reliability can be estimated by correlation between test trials and practice trials IAT scores (Greenwald et al, 2003). The basic idea is that the practice trials are the parallel form of test trials. However, the parallel assumption may be false. If it is true, the practice trials should be combined to test trials for creating a more reliable IAT score. In addition, this score do not estimate the internal consistency within test trials.

Because the first method does not create a reliability estimate equivalent to the other methods and item correlations are unknown for equating, therefore, the first method will not include in the main analysis. Although internal consistency coefficients based on split-half reliability, coefficient alpha, or factor analysis approaches produce different reliability estimates, the biases are not large because they are based on item covariances. Therefore, I will include in the same analysis. The correlation between test and practice trails has the different interpretation from other internal consistency coefficients. However, I will include in the analysis because the magnitude of correlation is equivalence to other internal consistency coefficients.

Factors Related to Internal Consistency of IAT

Although IAT has higher reliability than other implicit measures, it does not mean that every IAT is reliable. IATs which measure the same construct may have different psychometric properties. For example, IAT of stereotypes stimuli may be pictures or names of ethnic groups. Therefore, I will review several factors which may associate with reliability of IAT and make hypotheses for this study, if possible.

Group Homogeneity

Reliability is based on covariances between items. Covariances are the product of correlation and standard deviation. The correlation coefficient and standard deviation of scores are usually greater when groups are more heterogeneity. Therefore, participants who are heterogeneous will produce the greater reliability than homogeneous group of participants. In this study, I will compare reliability from college student samples with other samples. Because other samples have more chance to collect various participant characteristics, my hypothesis is that undergraduate participants give a lower reliability estimate.

Number of Trials

The number of trails is the number of stimuli used in step 4 or 7, which can be viewed as test length in psychometric textbooks. When test length increases, reliability of test increases. For IAT, the relationship between test length and reliability can be viewed in two aspects: number of difference scores and variability of IAT scores. First, if adding test trials by adding new paralleled stimuli to target or attribute category, the number of difference scores will increase in computing coefficient alpha of difference scores. The relationship between number of difference scores and reliability was illustrated by Spearman-Brown formula. Second, if adding old stimuli to test trials, the number of difference scores based on each stimulus does not change. However, because the difference score from each stimulus was combined by more than one pair, the variability of difference score will increase, as well as covariances between difference scores from different stimuli. The reliability coefficient will increase. Therefore, my hypothesis is that the more the number of trials, the more the reliability estimate.

The reliability of IAT and single target IAT may be different because of different number of trials also. IAT has four types of categories to compare reaction times: two targets and two attribute. However, single target IAT has only three types of categories: one target and two attribute. IAT have more number of difference scores than single target IAT; therefore, my hypothesis is that IAT have a greater reliability estimate than single target IAT.

Number of Stimuli Represented Each Category

Although the number of trials is equal, sometimes, the same stimulus may be represent more than one time. In other words, the number of stimuli represented in each category may be varied. If the number of stimuli represented in one category appears in IAT task twice or more, the response time from the same stimulus will be correlated from each other. Then, the item correlations will be greater and the reliability will be higher. However, reducing the number of stimulus represented each category may harm construct validity. Instead of responding based on superordinate category, test examiners may response based on feature recognition. This logic can apply for both number of stimuli for target and attribute category.

Nosek, Greenwald, and Banaji (2005) examined the relationship between number of stimuli represented each category and reliability based on correlation between the initial twenty trials and the last forty trials IAT scores. The results found that the correlations were not affected by number of stimuli except there is only one stimulus represent a category. However, as I said before, this correlation may not tell a true picture of internal consistency because two parts scores maybe not paralleled. Therefore, this study will examine the effect of the number of stimuli represented in each category for both targets and attributes.

Oppositeness of Attitude Targets

IAT is based on the reaction times between congruent and incongruent trials by comparing two attitude targets. The degree of oppositeness of two attitude targets affects interpretability of IAT score. The more the oppositeness, the clearer the interpretability of IAT scores as unidimensional scale. However, some studies compared one attitude target with neutral attitude target. For example, Ronay and Kim (2006) used the risk behaviors compared with square brackets. Nosek, et al. (2007) argued that it is difficult to find purely neutral attitude targets which can apply for everyone. Therefore, the IAT score may not purely represent the attitude of the desired target. In addition to interpretability, the oppositeness make the difference between reaction time between step 4 and step 7 higher, which imply that the variability of difference scores are higher. Therefore, I hypothesize that reliability of opposite targets is greater than non-opposite target.

Oppositeness of Attribute Categories

Sometimes, the attribute category was compared with the neutral categories. For example, Wiers, van de Luitgaarden, van den Wildenberg, & Smulders (2005) used IAT to find alcohol expectancies by finding contingency between alcohol and either negative or positive attributes. They argued that the heavy drinkers were not necessary to hold only positive expectancies. They may hold both positive and negative expectancies at the same time. Therefore, they used alcohol vs. soda drinks pairs with attributes of either positive or negative words coupled with neutral words. According to the reason provided in oppositeness of target categories, I hypothesize that opposite attributes are more reliable.

Types of IAT Domains

Both IAT and single target IAT are applied by various domains, such as interracial bias, personality, self-esteem, marketing, and politics. The IAT domains affect both correlation between implicit and explicit attitudes (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005) and criterion related validity (Greenwald, et al., in press). For example, Greenwald, et al. (in press) predicted criterions by both self-report attitude measures and IAT. IAT predicted consumer preferences in lower magnitude than self-report measures; however, they found the opposite direction in racial stereotype. Even if the differences may come from domains studied itself, such as social sensitivity of attitude domains, the differences may come from the different quality of IAT measures. Therefore, IAT domains are also included in this study.

Type of Target Stimuli

Types of target stimuli, such as names or faces, may provide different results. Type of stimuli may change attitude target although the stimuli come from the same domain. For example, Lane, et al. (2007) explained that, when faces of different ethnicity were shown, IAT will measure attitude toward people from different ethnicity. However, when the pictures of cities or buildings from different countries were shown, IAT measured attitude toward different cultures. In addition, IAT which desired to measure the same targets but different types of target stimuli may make the results different. For example, IAT measured implicit stereotype may use pronouns or idiographic information to represent me vs. others in implicit self-esteem. Hofmann, et al. (2005) showed that using idiographic information provided a greater correlation between IAT and self-report attitude measures than using pronouns. Therefore, I include type of target stimuli in this study.

Type of Attribute Stimuli

Attribute stimuli are not necessary to be positive and negative words. Sometimes, attribute stimuli are pictures or thematic words, such as which represent personality. Hofmann, et al. (2005) found that the types of attribute stimuli affect the correlation between IAT and explicit measures. The positive and negative nouns (e.g. peace vs. war) provided the highest correlation compared to positive and negative adjectives (e.g. good vs. bad), and thematic words reflecting different personality or stereotype. Thus, it is possible that reliabilities from different attribute types are different. Therefore, type of attribute stimuli is included in this study.

Experiment-based and web-based

The data from experiments and websites are different. Web-based participants are more heterogeneity than experiment-based participants, which most of them recruited from undergraduate students. From this fact, the web-based data may provide a higher reliability coefficient. However, data from web-based participants are prone to random error. For example, web-based participants may not follow IAT protocol or be distracted by environment. As the error variance is higher, the reliability of web-based data is lower. Therefore, I will explore reliability of both in this study.

Method

Literature Search

I used research studies which had been meta-analyzed by Greenwald, et al. (in press). These studied included IAT measures and reported predictive validity correlations. Therefore, some previous studies, which are not appropriate for their objective, are not included in my study also, such as which were interested in correlation between IAT and self-report measures only. Also, some recent studies, which were newer than 2007, are not included. They searched from research articles, conference papers, and unpublished studies. They also provided the collection of electronic articles in the first author website. Therefore, I focused on this pool of 122 studies from both published and unpublished reports initially. Thirty-eight studies reported reliability coefficients, which provided 63 independent samples and 111 reliability estimates.

Treatment of Internet Studies

There was only one internet-based (Friese, Bluemke, & Wanke, 2007) study which had a sample size of 1,548. The sample size was much larger than other studies, which may produce bias in weighted statistics. Therefore, I change the sample size of this study to 316, which is equal to the maximum sample size of the remaining studies. Because there is only one study, I will not use experiment vs. internet studies in moderation analysis.

Coding of Study Characteristics

These studies were coded to various factors which mentioned above. If the information given did not allow for definite coding, the data are coded as missing.

Characteristics of participants. Participants' characteristics from each study were coded as undergraduates or other samples, such as high school students, clinical samples, internet-based samples.

Characteristics of IAT. IAT characteristics are classified by many variables. First, the type of IAT was coded as IAT, single target IAT and Go/No-go IAT. Second, the domain of IAT was classified differently from Greenwald et al (in press) and Hofmann, et al. (2005). Because the number of studies reported reliability is not large, I classified domains of IAT to only four categories: intergroup/interpersonal (e.g. racial stereotype, gender stereotype, or interpersonal relationship), self/personality (e.g. self esteem or personality), clinic (e.g. drug abuse or phobia), and others (e.g. marketing or politics).

Third, target stimuli types are classified as names (e.g. names used frequently in each ethnic group), thematic words (e.g. risk behaviors), pictures (e.g. pictures of brand products), and pronouns (e.g. I, me, or them). Fourth, the attribute stimuli types are classified as valence nouns/adjective (e.g. positive or negative words), thematic words (e.g. words represented anxious and calm), and pictures (e.g. positive or negative pictures). Unlike Hoffman et al (2005), the valence nouns and adjectives are combined to the same category because a lot of study used both groups as attributes in IAT. Next, target oppositeness is classified as really opposite, which means that two attributes or category can be classified as two mutually exclusive category only (e.g. me and others), and somewhat opposite or neutral (e.g. Blacks vs. Whites or risks vs. square brackets). Also, attribute oppositeness is classified as really opposite (e.g. positive vs. negative words), and somewhat opposite or neutral (e.g. positive vs. neutral words).

The number of practice trails and test trials are also recorded. If there is no practice trial, it will be coded as 0 practice trial. The number of stimuli used for each target and attribute was also recorded. The coded number represents number of stimuli in one category only, such as a code is 5 when an IAT used 5 Blacks faces and 5 White faces. If the number of each category is not equal, such as 20 for positive words and 21 for negative words, the average was used.

Meta-analysis Procedure

Combination of multiple reliability reports. Because some research articles did multiple studies and reported multiple IAT reliability coefficients, this study uses independent samples as unit of analysis. However, some independent samples used more than one type of IAT. I treat reliability from both measures as a different unit. Five independent samples were reported both reliability estimates of standard IAT and single target IAT. One independent sample reported both reliability estimates of standard IAT and Go/No-go IAT. Therefore, there are 67 units of analysis from 63 independent samples. The average reliability reported in each unit of analysis is 1.71. The average of reliability coefficients from each unit is used in the main analysis.

Reliability coefficients. If the studies reported as split-half reliability, coefficient alpha from item parcels, or coefficient alpha from each different score, the reliability coefficients were used directly. However, some studies did not clearly report whether they use Spearman Brown Formula in split-half reliability. Three out of six reported the using of split-half reliability but not clearly stated the using of Spearman Brown Formula. I assume that they used this correction of test length. One study reported that they used correlation between halves only; I use Spearman-Brown formula to make the coefficient equivalent to other studies. Moreover, one study used confirmatory factor analysis from item parcels (Ashburn-Nardo, Knowles, & Monteith, 2003) and revealed the congeneric measurement model, I estimate reliability from factor loading and error variance (Lattin, et al., 2003) by formula 2.

If the studies reported the correlation between test and practice trials, the correlation is coded directly also. One study report confirmatory factor analysis result when latent factor consisted of score of test and practice trials (Ames, et al., 2007). The product of factor loadings from two scores is used to estimate the correlation between test and practice trials (Crocker &

Algina, 1986). However, if the studies reported reliability of difference scores, I will summarize in Appendix B but not included in the main analysis.

The summary of 67 reliability coefficient from all units of analysis was shown in Table 2. There are some studies that cannot include in main analysis. First, as the last row of the Table 2, 10 units were excluded because they reported coefficient alpha but did not report how to compute these coefficients. Next, one study was excluded because it reported the reliability estimate from correlation between error rates and reaction times (Ziegert and Hanget 2005), which not equivalent to other studies. Two reliability values from one study were excluded because of using reliability of difference scores formula (Cunningham, et al., 2001). Two split-half reliability were also excluded because they are reliability from Go/No-go IAT tasks (Teachman, 2007) which were not equivalent to standard IAT tasks. Finally, three correlations between test and practice trials were excluded because these studies did not report that they have practice trials. Therefore, the number of reliability coefficients used in the main analysis is 49.

Meta-analytic Computations. In this study, I use standard and weighted least squared general linear model, analyzed by SPSS. The weight variable is sample size of each independent variable. The reason why used both unweighted and weighted statistics is to determine the effect of sample size. The weighted statistics put more weight to units which contain larger number of participants. However, some units have very large sample size which may impact the results. Therefore, unweighted statistics which treated each unit of analysis equally, regardless of sample size, is also analyzed. If both ways of analysis provide the significant results, the results will be reliable.

I use this method because of time limit although I realize that this method was not the appropriate way. The more appropriate method is transformed reliability to more appropriate metrics (Rodrigues & Maeda, 2006).

Result

The histogram and boxplot of reliability estimates from 49 units are shown in Figure 1. The unweighted and weighted average are .77 (*SD* = .09) and .76 (*SD* = .96), respectively, based on overall sample size of 4,601 (corrected for the internet study). The reliability distribution is negatively skewed. Based on the boxplot, there is no outlier. Therefore, I kept all coefficients for further analysis.

Types of Reliability Coefficients

The reliability coefficients provided significantly different reliability coefficients, as shown in unweighted statistics, *F*(3, 45) = 11.35, *p* < .001, and weighted statistics, *F*(3, 45) = 16.96, *p* < .001. The unweighted and weighted averages and standard deviations of each type are provided in Table 2, as well as the Tukey's post hoc test results. From both statistics, the coefficient alpha from item parcels and difference scores produced the higher estimates than other methods. Also, the correlation between practice and test trials revealed the lower estimates than other methods.

Characteristics of Participants

As shown in Table 3, weighted statistics revealed that undergraduates provided a significantly higher reliability coefficient than other kinds of participants (*p* < .05). However, the unweighted statistics provided the opposite direction. Although the difference was not significant (*p* > .05), the unweighted mean of other kinds of participants was greater than undergraduates.

Characteristics of IAT type

The characteristics of IAT type which considered in this study are IAT type, domain of IAT, type of target and attribute stimuli, oppositeness of target and attribute stimuli, number of practice and test trials, and number of stimuli used in target and attribute categories. The moderator effects of categorical variables are shown in Table 3 and the effects of scale variables effects are shown in Table 4.

First, standard IAT and single target IAT have no significantly difference in reliability coefficients in unweighted statistics. However, analyzed by weighted statistics, the standard IAT has a marginally significantly greater reliability estimate than single category IAT ($p < .10$). Next, the effect of domain of IAT is not significant in unweighted statistics, but significant in weighted statistics. From the Tukey's post hoc comparison, the self/personality domain has a marginally significantly greater reliability than the clinical domain.

The type of target stimuli has a significant effect on reliability estimates also ($p < .05$ in unweighted mean and $p < .10$ in weighted mean). Thematic words have a significantly greater reliability estimate than names stimuli on both types of statistics ($p < .05$). Pictures stimuli have a significantly greater reliability coefficient than names stimuli on unweighted statistics only ($p < .05$). However, the type of attribute stimuli, target category oppositeness, and attribute category oppositeness do not have significant effect on reliability estimates in both types of statistics.

Finally, as shown in Table 4, the number of practice trials and test trials did not significantly correlated with reliability coefficients from both unweighted and weighted statistics. Also, the unweighted and weighted statistics reveal nonsignificant relationships between number of stimuli for either target or attribute and reliability estimates.

Discussion

This study provided the quantitative review of reliability coefficient of IAT. I summarized procedures for estimating reliability of IAT. Study characteristics which possibly related to reliability were also examined.

The result found that IAT had a mean population reliability of .75, which can be considered as quite high reliability. However, the different procedures for estimating reliability provided the different reliability estimates. The coefficient alpha from item parcels or each different score provided the highest estimates of .82. The split-half reliability provided the lower reliability and correlation between test and practice trials provided the lowest estimates. The possible reason why the split-half reliability was lower than coefficient alpha was that researchers did not use correction of test length. The result supported that correlation between test and practice trials were not equivalent to internal consistency estimates. However, the correlation of .65 may show that the test and practice trials are parallel measures.

The results revealed that participant characteristics effect was inconclusive. I will discuss this issue in limitation of this study. Regarding to IAT characteristics, standard IAT has a greater reliability than single target IAT. As predicted in number of trials factor, IAT provided the higher estimates because IAT have more number of reaction times to use as difference scores.

Next, the IAT studied in clinical setting produced the lowest estimates of reliability while self/personality domain provided the highest reliability. This difference may be explained by homogeneity of target stimuli (Cunningham, et al., 2001). The clinical setting target stimuli were heterogeneity (Houben & Wiers, 2006a). For example, most of IATs in clinical setting research are alcohol-related. Alcohol was a superordinate category which people felt toward types of alcohol differently. People may feel positive toward wine, but negative toward beer. However, the self/personality target stimuli were homogeneity. Most of IAT in this area are me/others targets (Bosson, Swann, & Pennebaker, 2000). People do not feel differently toward different pronouns of me and others. Therefore, the IAT of self/personality provided the higher reliability.

The type of target stimuli affected reliability estimates also. Thematic words and picture stimuli provided the greater reliability estimate than names stimuli. The possible explanation is that pictures and thematic words were classified easier than names. For example, pictures of people from each ethnic group were classified easily, compared to names from each group. The reaction times on names will vary unsystematically than picture stimuli. Therefore, reliability toward names is lower than other types of target stimuli. On the other hand, the type of attribute stimuli was not significant. In spite of nonsignificant effect, the pictures stimuli showed the largest reliability. However, only four units used picture attributes. Therefore, it is possible that this study is lack of power in detecting this difference.

Although the effect of target oppositeness was not significant, the direction of relationship was the same as prediction that the opposite target had a larger reliability estimate than somewhat opposite target. The larger pool of studies is required to gain a better estimate of the effect of target oppositeness. Also, I cannot conclude the effect of attribute oppositeness because the number of studies provided the valence compared with neutral stimuli IAT was only two.

The number of practice trials was not significant, which can explain in various ways. In addition to the low power of this study, the number of practice of trials was confounded with method of reliability used. The correlation between practice and test trials, which was lower than other methods, can be calculated only IAT with practice and test trials. The number of test trials was not significant also. However, the number of test trials effect was in the same direction as prediction. The larger pool of study is required to detect this effect also. The number of stimuli for each target and attribute did not have significant effect on reliability estimates. Although the number of stimuli effect was not significant, the number of stimuli for each attribute tended to have negative effect on reliability.

Limitation

Although this study provided background about reliability of IAT, this study is far from the good meta-analysis article. The first limitation is the pool of studies. As I used the pool of studies from Greenwald, et al. (in press), many studies were not included in my review. This limitation caused this study have a low power in detecting effect and generalizability of this study. In addition, many reliability coefficients are not included in my analysis because of lack of information, such as how to estimate coefficient alpha. If I contacted authors of dropped studies, this study will include a larger number of coefficients and has more generalizability. The more accurate reliability estimates will be included in this study also, such as the information whether Spearman-Brown formula was used in their studies. The next limitation is the method used in data analysis. The better statistical model was provided in Rodrigues and Maeda (2006). The result of this study was only a brief estimate of reliability parameter and moderation effects. The revised version of this study is required to provide a better conclusion. Finally, this study did not test the effect of all factors at once, like multiple regression analysis. Therefore, the moderation effect did not consider the correlation between factors. For example, the self/personality IAT usually have pronoun targets. The effects of two variables are confounded.

Suggestion for IAT Research

Although IAT provide the largest reliability, different IAT measures give the different reliability, range from .50 to .90. Therefore, I recommended researchers to give the reliability in every study, instead of assuming that IAT have a better reliability than other implicit measures. Moreover, the method used for estimating reliability should be standardized because readers can understand the reliability of IAT directly and compare with other studies easily. I recommend researchers to use coefficient alpha of difference scores. It provides the high reliability, according to the result of this study. It represents internal consistency of IAT measures and also solves the problem of subjectivity in splitting item parcels.

As another suggestion, although single category IAT had a lower reliability estimate, it does not mean that researchers should not use this type of IAT. However, the single target IAT has the benefits of interpretation the attitude of one target without comparing with other category. Therefore, researchers should balance between advantage of interpretability and disadvantage of reliability reduction.

Reference

* Indicates studies provided reliability estimates

** Indicates studied included in the meta-analysis

**Ames, S. L., Grenard, J. L., Thush, C., Sussman, S., Wiers, R. W., & Stacy, A. W. (2007). Comparison of indirect assessments of association as predictors of marijuana use among at-risk adolescents. *Experimental and Clinical Psychopharmacology, 15*, 204-218.

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

**Ashburn-Nardo, L., Knowles, M. L., & Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition, 21*, 61-87.

**Banse, R. (2007). Implicit attitudes towards romantic partners and ex-partners: A test of the reliability and validity of the IAT. *International Journal of Psychology, 42*, 149-157.

*Banse, R., & Fischer, I. (2002). *Implicit and explicit aggressiveness and the prediction of aggressive behaviour*. Paper presented at the 11th conference on personality of the European Association of Personality Psychology, Jena.

*Banse, R., Grune, J., & Kreft, V. (2002). *Implicit attitudes towards romantic partners: Convergent and discriminant validity*. Paper presented at the 13th General meeting of the European Association of Experimental Social Psychology, San Sebastian.

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

**Brunstein, J. C., & Schmitt, C. H. (2004). Assessing individual differences in achievement motivation with the Implicit Association Test. *Journal of Research in Personality, 38*, 536-555.

*Carney, D. R. (2006). *The faces of prejudice: On the malleability of the attitude-behavior link*.Unpublished manuscript.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Thomson.

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

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

**Ellwart, T., Rinck, M., & Becker, E. S. (2006). From fear to love: Individual differences in implicit spider associations. *Emotion, 6*, 18-27.

**Friese, M., Bluemke, M., & Wanke, M. (2007). Predicting voter behavior with implicit attitude measures: The 2002 German parliamentary election. *Experimental Psychology, 54*, 247-255.

**Friese, M., Hofmann, W., & Wanke, M. (2008). When impulses takes over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology, 47*, 397-419.

**Gabriel, U., Banse, R., & Hug, F. (2007). Predicting private and public helping behavior by implicit prejudice and the motivation to control prejudiced reactions. *British Journal of Social Psychology, 46*, 365-382.

**Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: Individual differences in associative strength moderate category based and individuating impression formation. *Journal of Experimental Social Psychology, 39*, 16-30.

**Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology, 33*, 573-589.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (in press). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*.

**Hofmann, W., & Friese, M. (2008). Impulses got the better of me: Alcohol moderates the influence of implicit attitudes toward food cues on eating behavior. *Journal of Abnormal Psychology, 117*, 420-427.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369-1385.

**Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations, 11*, 69-87.

*Houben, K., & Wiers, R. W. (2006a). Assessing implicit alcohol associations with the Implicit Association Test: Fact or artifact? *Addictive Behaviors, 31*, 1346-1362.

*Houben, K., & Wiers, R. W. (2006b). A test of the salience asymmetry interpretation of the alcohol-IAT. *Experimental Psychology, 53*, 292-300.

**Karpinski, A., & Steinman, R. B. (2006). The single category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*, 16-32.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV. What we know (so far) about the method. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59-102). New York: Guilford.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Thomson.

**Levesque, C., & Brown, K. W. (2004). *Overriding motivational automaticity: Mindfulness as a moderatorr of the influence of implicit motivation on day-to-day behavior*.Unpublished manuscript.

**Marsh, K. L., Johnson, B. T., & Scott-Sheldon, L. A. J. (2001). Heart versus reason in condom use: Implicit vs. explicit attitudinal predictors of sexual behavior. *Zeitschrift fur Experimentelle Psychologie, 48,* 161-175.

**Mauss, I. B., Evers, C., Wilhelm, F. H., & Gross, J. J. (2006). How to bite your tongue without blowing your top: Implicit evaluation of emotion regulation predicts affective responding to anger provocation. *Personality and Social Psychology Bulletin, 32*, 589-602.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265-292). New York: Psychology Press.

**Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology, 44*, 29-45.

*Plessner, H., Haar, T., Hoffman, K., Stark, R., & Wanke, M. (2006). *Directness of attitude measure and speed of information processing as constituents of consumer's attitude behavior correspondence*.Unpublished manuscript.

**Robinson, M. D., Mitchell, K. A., Kirkeby, B. S., & Meier, B. P. (2006). The self as a container: Implications for implicit self-esteem and somatic symptoms. *Metaphor and Symbol, 21*, 147-167.

Rodrigues, M., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322.

*Ronay, R., & Kim, D. (2006). Gender differences in explicit and implicit risk attitudes: A socially facilitated phenomenon. *British Journal of Social Psychology, 45*, 397-419.

**Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations, 10*, 359-372.

**Schnabel, K., Banse, R., & Asendorpf, J. B. (2006a). Assessment of implicit personality self-concept using the Implicit Association Test (IAT): Concurrent assessment of anxiousness and angriness. *British Journal of Social Psychology, 45*, 373-396.

**Schnabel, K., Banse, R., & Asendorpf, J. B. (2006b). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology, 53*, 69-76.

**Steffens, M. C., & Konig, S. S. (2006). Predicting spontaneous Big Five behavior with Implicit Association Tests. *European Journal of Psychological Assessment, 22*, 13-20.

*Teachman, B. A. (2007). Evaluating implicit spider fear associations using the Go/No-go Association Task. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 157-167.

*Thush, C., & Wiers, R. W. (2007). Explicit and implicit alcohol-related cognitions and the prediction of future drinking in adolescents. *Addictive Behaviors, 32*, 1367-1383.

*van den Wildenberg, E., Beckers, M., van Lambaart, F., Conrod, P. J., & Wiers, R. W. (2006). Is the strength of implicit alcohol associations correlated with alcohol-induced heart-rate acceleration? *Alcoholism: Clinical and Experimental Research, 30*, 1336-1348.

*Wiers, R. W., Houben, K., & de Kraker, J. (2007). Implicit cocaine associations in active cocaine users and controls. *Addictive Behaviors, 32*, 1284-1289.

**Wiers, R. W., van de Luitgaarden, J., van den Wildenberg, E., & Smulders, F. T. Y. (2005). Challenging implicit and explicit alcohol-related cognitions in young heavy drinkers. *Addiction, 100*, 806-819.

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

*Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology, 90*, 554-562.

Appendix A

Formula of Reliability from Reaction Time Differences of Different Stimuli

The variances of true score and observed score of difference score in each stimulus are

$$\sigma^2_{T_3-T_5} = \sigma^2_{T_3} + \sigma^2_{T_5} - \sigma_{T_3T_5} \qquad \text{(A1)}$$

$$\sigma^2_{O_3-O_5} = \sigma^2_{O_3} + \sigma^2_{O_5} - \sigma_{O_3O_5} \qquad \text{(A2)}$$

The general formula of reliability of composite scores is

$$\rho_{CC'} = \frac{\sigma^2_{T_C}}{\sigma^2_C} \qquad \text{(A3)}$$

The composite scores are the sum of difference scores from stimuli

$$T_C = \sum_{i=1}^{k}\left(T_{4_i} - T_{7_i}\right)$$

$$C = \sum_{i=1}^{k}\left(O_{4_i} - O_{7_i}\right)$$

The variances of true and observed composite scores are

$$\sigma^2_{T_C} = \textstyle\sum_{i=1}^{k} \sigma^2_{(T_4-T_7)_i} + \sum_{i\neq j} \sigma_{(T_4-T_7)_{ij}} \qquad \text{(A4)}$$

$$\sigma^2_C = \textstyle\sum_{i=1}^{k} \sigma^2_{(O_4-O_7)_i} + \sum_{i\neq j} \sigma_{(O_4-O_7)_{ij}} \qquad \text{(A5)}$$

The first term of right hand size of equation A4 and A5 can be expanded as

$$\sum_{i=1}^{k} \sigma^2_{(T_3-T_5)_i} = \sum_{i=1}^{k}\left(\sigma^2_{T_3} + \sigma^2_{T_5} - \sigma_{T_3T_5}\right) = \sum_{i=1}^{k} \sigma^2_{(T_3)_i} + \sum_{i=1}^{k} \sigma^2_{(T_5)_i} - \sum_{i=1}^{k} \sigma_{(T_3)_i(T_5)_i}$$

$$\sum_{i=1}^{k} \sigma^2_{(O_3-O_5)_i} = \sum_{i=1}^{k}\left(\sigma^2_{O_3} + \sigma^2_{O_5} - \sigma_{O_3O_5}\right) = \sum_{i=1}^{k} \sigma^2_{(O_3)_i} + \sum_{i=1}^{k} \sigma^2_{(O_5)_i} - \sum_{i=1}^{k} \sigma_{(O_3)_i(O_5)_i}$$

The second term of right hand size of equation A4 and A5 can be expanded as

$$\sum_{i\neq j}\sigma_{(T_3-T_5)_{ij}} = \sum_{i\neq j}\left(\sigma_{(T_3)_{ij}} + \sigma_{(T_5)_{ij}} - \sigma_{(T_3)_i(T_5)_j} - \sigma_{(T_3)_j(T_5)_i}\right)$$

$$= \sum_{i\neq j}\left(\sigma_{(T_3)_{ij}} + \sigma_{(T_5)_{ij}} - 2\sigma_{(T_3)_i(T_5)_j}\right) = \sum_{i\neq j}\sigma_{(T_3)_{ij}} + \sum_{i\neq j}\sigma_{(T_5)_{ij}} - 2\sum_{i\neq j}\sigma_{(T_3)_i(T_5)_j}$$

With the same step, the second term of right hand size of equation A5 is

$$\sum_{i\neq j}\sigma_{(O_4-O_7)_{ij}} = \sum_{i\neq j}\sigma_{(O_4)_{ij}} + \sum_{i\neq j}\sigma_{(O_7)_{ij}} - 2\sum_{i\neq j}\sigma_{(O_4)_i(O_7)_j}$$

Replace the first and second term in the right hand size of equation A4

$$\sigma^2_{T_C} = \sum_{i=1}^{k}\sigma^2_{(T_3)_i} + \sum_{i=1}^{k}\sigma^2_{(T_5)_i} - \sum_{i=1}^{k}\sigma_{(T_3)_i(T_5)_i} + \sum_{i\neq j}\sigma_{(T_3)_{ij}} + \sum_{i\neq j}\sigma_{(T_5)_{ij}} - 2\sum_{i\neq j}\sigma_{(T_3)_i(T_5)_j}$$

$$\sigma^2_{T_C} = \left(\sum_{i=1}^{k}\sigma^2_{(T_3)_i} + \sum_{i\neq j}\sigma_{(T_3)_{ij}}\right) + \left(\sum_{i=1}^{k}\sigma^2_{(T_5)_i} + \sum_{i\neq j}\sigma_{(T_5)_{ij}}\right) - \left(\sum_{i=1}^{k}\sigma_{(T_3)_i(T_5)_i} + 2\sum_{i\neq j}\sigma_{(T_3)_i(T_5)_j}\right)$$

$$\sigma^2_{T_C} = \sigma^2_{T_3} + \sigma^2_{T_5} - 2\left(\sum_{i=1}^{k}\sigma_{(T_3)_i(T_5)_i} + \sum_{i\neq j}\sigma_{(T_3)_i(T_5)_j}\right) + \sum_{i=1}^{k}\sigma_{(T_3)_i(T_5)_i}$$

$$\sigma^2_{T_C} = \sigma^2_{T_3} + \sigma^2_{T_5} - 2\left(\sum_{i=1}^{k}\sigma_{(O_3)_i(O_5)_i} + \sum_{i\neq j}\sigma_{(O_3)_i(O_5)_j}\right) + \sum_{i=1}^{k}\sigma_{(O_3)_i(O_5)_i}$$

$$\sigma^2_{T_C} = \rho_{33}\sigma^2_3 + \rho_{55}\sigma^2_5 - 2\sigma_{35} + \sum_{i=1}^{k}\sigma_{(O_3)_i(O_5)_i}$$

$$\sigma^2_{T_C} = \rho_{33}\sigma^2_3 + \rho_{55}\sigma^2_5 - 2\rho_{35}\sigma_3\sigma_5 + \textstyle\sum_{i=1}^{k}\rho_{35_i}\sigma_{3_i}\sigma_{5_i} \qquad \text{(A6)}$$

Replace the first and second term in the right hand size of equation A5, the result is

$$\sigma^2_C = \sigma^2_3 + \sigma^2_5 - 2\rho_{35}\sigma_3\sigma_5 + \textstyle\sum_{i=1}^{k}\rho_{35_i}\sigma_{3_i}\sigma_{5_i} \qquad \text{(A7)}$$

Replace numerator and denominator of equation A3 by equation A6 and A7, the result is

$$\rho_{CC'} = \frac{\sigma^2_{T_C}}{\sigma^2_C} = \frac{\rho_{33}\sigma^2_3+\rho_{55}\sigma^2_5-2\rho_{35}\sigma_3\sigma_5+\sum_{i=1}^{k}\rho_{35_i}\sigma_{3_i}\sigma_{5_i}}{\sigma^2_3+\sigma^2_5-2\rho_{35}\sigma_3\sigma_5+\sum_{i=1}^{k}\rho_{35_i}\sigma_{3_i}\sigma_{5_i}} \qquad \text{(A8)}$$

Appendix B

Characteristics of the 63 independent samples reported reliability coefficients

Citation	Sample	N	NR	IAT-T	Domain	Targets	Attributes	Relia	Method
Ames et al. (2007)	1	212	2	IAT	Clinical	Marijuana vs. Other Pictures	Excited vs. Neutral	0.59	COR-PT
				IAT	Clinical	Marijuana vs. Other Pictures	Relaxed vs. Neutral	0.57	COR-PT
Asendorpf, Banse, & Mucke (2002)	1	139	2	IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.89	ALPHA-P-4
				IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.82	ALPHA-P-4
Ashburn-Nardo, Knowles, & Monteith (2003)	1	316	1	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.82	ALPHA-P-4
*Banse & Fischer (2002)	1	50	2	IAT	Self/Personality	Me vs. Others	Aggressive vs. Non-aggressive Interactions	0.86	N/A
				IAT	Self/Personality	Me vs. Others	Aggressive vs. Non-aggressive Traits	0.68	N/A
	2	44	2	IAT	Self/Personality	Me vs. Others	Aggressive vs. Non-aggressive Interactions	0.81	N/A
				IAT	Self/Personality	Me vs. Others	Aggressive vs. Non-aggressive Traits	0.79	N/A
Banse (2007)	3	21	1	IAT	Intergroup	Partner vs. Stranger	Positive vs. Negative Words	0.81	ALPHA-P-3
		46	1	IAT	Intergroup	Partner vs. Stranger	Positive vs. Negative Words	0.84	ALPHA-P-3
		50	1	IAT	Intergroup	Partner vs. Stranger	Positive vs. Negative Words	0.8	ALPHA-P-3
		19	1	IAT	Intergroup	Partner vs. Stranger	Positive vs. Negative Words	0.89	ALPHA-P-3
*Banse, Grune, & Kreft (2002)	1	96	1	IAT	Intergroup	Partner vs. Ideal Partner	Positive vs. Negative Words	0.9	N/A
Bosson, Swann, & Pennebaker (2000)	1	84	1	IAT	Self/Personality	Me vs. Others	Positive vs. Negative Words	0.88	ALPHA-D
Brunstein & Schmitt (2004)	1	88	1	IAT	Self/Personality	Me vs. Others	Successful vs. Not Successful	0.82	ALPHA-P-4
*Carney (2006)	1	62	1	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.42	N/A
Cunnighham, Preacher, & Banaji (2001)	1	99	2	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.78	RD
				IATG	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.63	RD
Egloff & Schmukle (2002)	1	41	2	IAT	Self/Personality	Me vs. Others	Anxiety vs. Calm	0.77	ALPHA-D
				IAT	Self/Personality	Me vs. Others	Anxiety vs. Calm	0.8	ALPHA-D
	2	20	1	IAT	Self/Personality	Me vs. Others	Anxiety vs. Calm	0.71	ALPHA-D
	3	20	1	IAT	Self/Personality	Me vs. Others	Anxiety vs. Calm	0.8	ALPHA-D
Ellwart, Rinck, & Becker (2006)	1	48	1	IAT	Clinical	Spider vs. Butterfly	Positive vs. Negative Words	0.84	SH
	2	18	1	IAT	Clinical	Spider vs. Butterfly	Positive vs. Negative Words	0.84	SH
Friese, Bluemke, & Wanke (2007)	1	1548	1	ST-IAT	Others	Political Parties	Positive vs. Negative Words	0.67	SH
Friese, Hofmann, & Wanke (2008)	1	88	1	IAT	Others	Fruit vs. Chocolate	Positive vs. Negative Words	0.93	ALPHA-P-3
	2	69	1	ST-IAT	Others	Chips	Positive vs. Negative Words	0.73	ALPHA-P-3
	3	48	1	ST-IAT	Clinical	Beer	Positive vs. Negative Words	0.81	ALPHA-P-3
Gabriel, Banse, & Hug (2007)	1	69	1	IAT	Intergroup	Homosexual vs. Heterosexual	Positive vs. Negative Words	0.78	ALPHA-P-3
Gawronski, Ehrenberg, Banse, Zukova, & Klauer (2003)	1	122	1	IAT	Intergroup	Men vs. Women	Career vs. Household	0.8	ALPHA-P-3
	2	60	1	IAT	Intergroup	Men vs. Women	Career vs. Household	0.75	ALPHA-P-3
Gawronski, Geschke, & Banse (2003)	1	70	1	IAT	Intergroup	German vs. Turkish	Positive vs. Negative Words	0.9	ALPHA-P-3
Hofman & Friese (2008)	1	63	1	ST-IAT	Others	M & M	Positive vs. Negative Pictures	0.83	ALPHA-P-4
Hofmann & Gschwender, Castelli, & Schmitt (2008)	1	86	2	IAT	Intergroup	African vs. Italian	Positive vs. Negative Words	0.9	ALPHA-P-4
				IAT	Clinical	Flowers vs. Insects	Positive vs. Negative Words	0.88	ALPHA-P-4
Houben & Wiers (2006a)	1	96	4	IAT	Clinical	(Alcohol or Beer) vs. (Soda or Animals)	Positive vs. Neutral	0.46	COR-PT
				IAT	Clinical	(Alcohol or Beer) vs. (Soda or Animals)	Negative vs. Neutral	0.44	COR-PT

Citation	Sample	N	NR	IAT-T	Domain	Targets	Attributes	Relia	Method
				IAT	Clinical	(Alcohol or Beer) vs. (Soda or Animals)	Arousal vs. Neutral	0.52	COR-PT
				IAT	Clinical	(Alcohol or Beer) vs. (Soda or Animals)	Sedation vs. Neutral	0.46	COR-PT
Houben & Wiers (2006b)	1	46	2	IAT	Clinical	Familiar Alcohol vs. Unfamiliar Soft Drinks	Positive vs. Negative Words	0.84	COR-PT
				IAT	Clinical	Unfamiliar Alcohol vs. Familiar Soft Drinks	Positive vs. Negative Words	0.87	COR-PT
Karpinski & Steinman (2006)	1	56	3	ST-IAT	Others	Coke	Positive vs. Negative Words	0.61	SH-3
				ST-IAT	Others	Pepsi	Positive vs. Negative Words	0.69	SH-3
				IAT	Others	Coke vs. Pepsi	Positive vs. Negative Words	0.82	COR-PT
	2	66	2	ST-IAT	Self/Personality	Me	Positive vs. Negative Words	0.73	SH-3
				IAT	Self/Personality	Me vs. Others	Positive vs. Negative Words	0.58	COR-PT
	3	118	3	ST-IAT	Intergroup	Whites	Positive vs. Negative Words	0.7	SH-3
				ST-IAT	Intergroup	Blacks	Positive vs. Negative Words	0.55	SH-3
				IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.75	COR-PT
	4	84	2	ST-IAT	Intergroup	Women	Positive vs. Negative Words	0.85	SH-3
				IAT	Intergroup	Men vs. Women	Positive vs. Negative Words	0.78	COR-PT
*Levesque & Brown (2004)	1	78	2	IAT	Self/Personality	Me vs. Others	Autonomy vs. Heteronomy	0.75	ALPHA-P-2
				IAT	Self/Personality	Autonomy vs. Heteronomy	Positive vs. Negative Words	0.67	ALPHA-P-2
	2	69	1	IAT	Self/Personality	Me vs. Others	Autonomy vs. Heteronomy	0.78	ALPHA-P-2
	3	78	1	IAT	Self/Personality	Me vs. Others	Autonomy vs. Heteronomy	0.8	ALPHA-P-2
Marsh, Johnson, & Scott-Sheldon (2001)	1	97	2	IAT	Others	Condom vs. Non-Condom	Positive vs. Negative Words	0.57	SH
				IAT	Others	Me vs. Others	Condom vs. Non-Condom	0.75	SH
Mauss, Evers, Wilhelm, & Gross (2006)	1	245	6	IAT	Self/Personality	Emotional Regulation vs. Emotional Expression	Positive vs. Negative Words	0.86	ALPHA-D
Perugini (2005)	1	50	1	IAT	Clinical	Smoking vs. Exercise	Positive vs. Negative Words	0.8	ALPHA-P-4
	2	113	1	IAT	Others	Snacks vs. Fruits	Positive vs. Negative Words	0.86	ALPHA-P-4
*Plessner, Haar, Hoffman, Stark, & Wanke (2006)	1	40	1	IAT	Others	Recycled Papers vs. White Papers	Positive vs. Negative Words	0.96	N/A
	2	112	3	IAT	Others	Newspaper 1 vs. Newspaper 2	Positive vs. Negative Words	0.91	N/A
				ST-IAT	Others	Newspaper 1	Positive vs. Negative Words	0.76	N/A
				ST-IAT	Others	Newspaper 2	Positive vs. Negative Words	0.76	N/A
Robinson, Mitchell, Kirkeby, & Meier (2006)	1	96	1	IAT	Self/Personality	Me vs. Others	Positive vs. Negative Words	0.81	SH
	2	61	1	IAT	Self/Personality	Me vs. Others	Positive vs. Negative Words	0.74	SH
Ronay & Kim (2006)	1	126	2	IAT	Self/Personality	Risk vs. [ ]	Gain vs. Loss	0.73	N/A
				IAT	Self/Personality	Risk behaviors vs. [ ]	Gain vs. Loss	0.95	N/A
Rudman & Ashmore (2007)	1	64	2	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.69	COR-PT
				IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Traits	0.71	COR-PT
	2	89	1	IAT	Intergroup	Jewish vs. Christian	Positive vs. Negative Traits	0.61	COR-PT
	3	89	2	IAT	Intergroup	White vs. Asians	Positive vs. Negative Words	0.59	COR-PT
				IAT	Intergroup	White vs. Asians	Positive vs. Negative Traits	0.6	COR-PT
	4	126	2	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.57	COR-PT
				IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Traits	0.63	COR-PT
Schnabel, Banse, Asendorpf (2006a)	1	300	4	IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.78	ALPHA-P-4

Citation	Sample	N	NR	IAT-T	Domain	Targets	Attributes	Relia	Method
				IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.76	ALPHA-P-4
				IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.83	ALPHA-P-4
				IAT	Self/Personality	Me vs. Others	Shy vs. Non-shy	0.77	ALPHA-P-4
Schnabel, Banse, Asendorpf (2006b)	1	100	2	IAT	Self/Personality	Me vs. Others	Anxious vs. Self-confident	0.72	ALPHA-P-2
				IAT	Self/Personality	Me vs. Others	Angry vs. Self-controlled	0.66	ALPHA-P-2
Steffens & Konig (2006)	1	89	5	IAT	Self/Personality	Me vs. Others	Emotional Stable vs. Emotional Labile	0.76	ALPHA-D
				IAT	Self/Personality	Me vs. Others	Extraverted vs. Introverted	0.51	ALPHA-D
				IAT	Self/Personality	Me vs. Others	Culturally Interested vs. Not Culturally Interested	0.68	ALPHA-D
				IAT	Self/Personality	Me vs. Others	Agreeable vs. Not Agreeable	0.7	ALPHA-D
				IAT	Self/Personality	Me vs. Others	Conscientious vs. Not Conscientious	0.81	ALPHA-D
Teachman (2007)	1	34	2	IATG	Clinical	Spider vs. Other Animals	Afraid vs. Calm	0.46	SH
				IATG	Clinical	Fire vs. Other Elements	Afraid vs. Calm	0.49	SH
Thush & Wiers (2007)	1	100	4	ST-IAT	Clinical	Alcohol	Positive vs. Neutral	0.51	N/A
				ST-IAT	Clinical	Alcohol	Negative vs. Neutral	0.52	N/A
				ST-IAT	Clinical	Alcohol	Arousal vs. Neutral	0.46	N/A
				ST-IAT	Clinical	Alcohol	Sedation vs. Neutral	0.43	N/A
Van den Wildenberg, Beckers, van Lambaart, Conrod, & Wiers (2006)	1	48	6	IAT	Clinical	Alcohol vs. Soda	Positive vs. Neutral	0.62	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Negative vs. Neutral	0.72	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Arousal vs. Neutral	0.76	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Sedation vs. Neutral	0.62	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Material vs. Neutral	0.4	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Approach vs. Avoidance	-0.01	COR-PT
Wiers, Houben, & de Kraker (2007)	1	32	4	IAT	Clinical	Cocaine vs. Sports	Positive vs. Neutral	0.45	N/A
				IAT	Clinical	Cocaine vs. Sports	Negative vs. Neutral	0.47	N/A
				IAT	Clinical	Cocaine vs. Sports	Arousal vs. Neutral	0.63	N/A
				IAT	Clinical	Cocaine vs. Sports	Sedation vs. Neutral	0.27	N/A
Wiers, van de Luitgaarden, van den Wildenberg, & Smulders (2005)	1	96	2	IAT	Clinical	Alcohol vs. Soda	Positive vs. Negative Words	0.65	COR-PT
				IAT	Clinical	Alcohol vs. Soda	Arousal vs. Sedation	0.68	COR-PT
Ziegert & Hanges (2005)	1	103	1	IAT	Intergroup	Whites vs. Blacks	Positive vs. Negative Words	0.63	COR-RTE

*=unpublished report; N = number of subjects in independent sample; NR = number of distinct reliability coefficient in independent sample; IAT-T = IAT Type; ST-IAT = single target IAT; IATG = IAT Go/NoGo Task; Relia = Reported Reliability Coefficient; COR-PT = Correlation between test and practice trials; ALPHA-P = Coefficient alpha of item parcels, the number attached is number of parcels; ALPHA-D = Coefficient alpha of difference scores; COR-RTE = Correlation between Reaction Time and Errors; RD = Reliability of Difference Scores

Table 1

Overall procedure of IAT and single target IAT

Step	Trails	Key	IAT	Single target IAT
1		Left	Whites	
		Right	Blacks	
2		Left	Positive	Positive
		Right	Negative	Negative
3	Practice	Left	Whites and Positive	Whites and Positive
		Right	Blacks and Negative	Negative
4	Test	Left	Whites and Positive	Whites and Positive
		Right	Blacks and Negative	Negative
5		Left	Blacks	
		Right	Whites	
6	Practice	Left	Blacks and Positive	Positive
		Right	Whites and Negative	Whites and Negative
7	Test	Left	Blacks and Positive	Positive
		Right	Whites and Negative	Whites and Negative

Note. If the combination tasks used all reaction times, the step 3 and 6 do not exist.

Table 2

Type of reliability reported in 67 studies and descriptive statistics of each type of reliability coefficient by both unweighted and weighted statistics

Type of Reliability	*N*	N_A	Unweighted		Weighted	
			M	*SD*	*M*	*SD*
Split-Half Reliability	12	10	$.74_{a, b}$	.09	$.71_a$	.79
Coefficient Alpha from Item Parcels	23	23	$.83_a$	.06	$.82_b$	.53
Coefficient Alpha from Difference Scores from Each Stimulus	6	6	$.78_a$	.07	$.82_b$	.70
Correlation between Practice and Test Trials	13	10	$.69_b$	.10	$.65_a$	.85
Others (correlation between reaction time and error and reliability of difference scores)	3	0				
Cannot be coded	10	0				
All studies	67	49	.77	.09	.76	.96

Note. N_A = number of units used in the meta-analysis. Subscripts denote comparisons within a column. Means with different subscripts are significantly different from one another.

Table 3

The descriptive statistics and one-way ANOVA, included both unweighted and weighted statistics, result of reliability coefficients when grouped studies characteristics are independent variables

Studies Characteristics	*N*	N_S	Unweighted		Weighted	
			M	*SD*	*M*	*SD*
Participants type			*F*(1,47) = 0.30, *p* = .59		*F*(1, 47) = 4.78, *p* = .034	
Undergraduates	42	3899	0.77	0.09	0.77	0.89
Others	7	702	0.79	0.12	0.69	1.17
IAT Type			*F*(1, 47) = 1.30, *p* = .26		*F*(1, 47) = 2.88, *p* = .096	
Standard IAT	41	3781	0.78	0.10	0.77	0.96
Single target IAT	8	820	0.74	0.09	0.71	0.81
Domain of IAT			*F*(3, 44) = 0.25, *p* = .86		*F*(3, 44) = 2.25, *p* = .095	
Intergroup/Interpersonal	17	1545	0.76	0.10	$0.75_{a, b}$	0.94
Self/Personality	17	1640	0.78	0.08	0.80_{a}	0.76
Clinic	6	472	0.76	0.11	0.68_{b}	1.04
Others	8	858	0.77	0.09	$0.75_{a, b}$	1.09
Target Stimuli Type			*F*(3, 43) = 3.63, *p* = .021		*F*(3, 43) = 2.68, *p* = .053	
Names	6	550	0.68_{b}	0.09	0.67_{b}	0.89
Thematic Words	10	808	0.82_{a}	0.06	0.81_{a}	0.60
Pictures	15	1435	0.79_{a}	0.10	$0.77_{a, b}$	1.11
Pronouns	16	1633	$0.77_{a, b}$	0.09	$0.76_{a, b}$	0.87
Attribute Stimuli Type			*F*(2, 40) = 0.73, *p* = .49		*F*(2, 40) = 0.96, *p* = .39	
Valence Nouns/Adjective	26	2445	0.79	0.09	0.78	0.85
Thematic Words	13	1338	0.77	0.10	0.75	1.07
Pictures	4	268	0.83	0.08	0.83	0.73
Target Category Oppositeness			*F*(1, 37) = 0.13, *p* = .72		*F*(1, 37) = 2.28, *p* = .14	
Really Opposite	20	1909	0.79	0.08	0.80	0.69
Somewhat Opposite or Neutral	19	1719	0.78	0.11	0.75	1.16
Attribute Category Oppositeness			*F*(1, 46) = 1.30, *p* = .26		*F*(1, 46) = 105, *p* = .31	
Really Opposite	46	3976	0.78	0.09	0.77	0.88
Somewhat Opposite or Neutral	2	528	0.70	0.17	0.72	2.70

Note. N_S = number of overall sample size in each group (Four hundred and thirty-six participants were double counted when they provided both IAT and single category IAT reliability coefficients). Subscripts denote comparisons within a column. Means with different subscripts are marginally significantly different from one another (*p* < .10).

Table 4

The descriptive statistics of trials and stimuli used in IAT and correlation of those with reliability coefficients when both unweighted and weighted statistics were used.

IAT Characteristics	*N*	Unweighted				Weighted			
		M	*SD*	*r*	*p*	*M*	*SD*	*r*	*p*
Number of Practice Trials	43	16.33	15.99	-.110	.24	13.89	143.83	-.030	.42
Number of Test Trials	43	76.37	48.43	.195	.11	66.41	370.43	.141	.18
Number of Stimuli for each target	39	6.90	3.50	.134	.21	6.25	24.44	.030	.43
Number of Stimuli for each attribute	39	8.15	5.21	-.147	.19	7.64	47.45	-.154	.17

Figure Captions

1. Histogram and boxplot of 49 reliabilities coefficients which were used in the main analysis

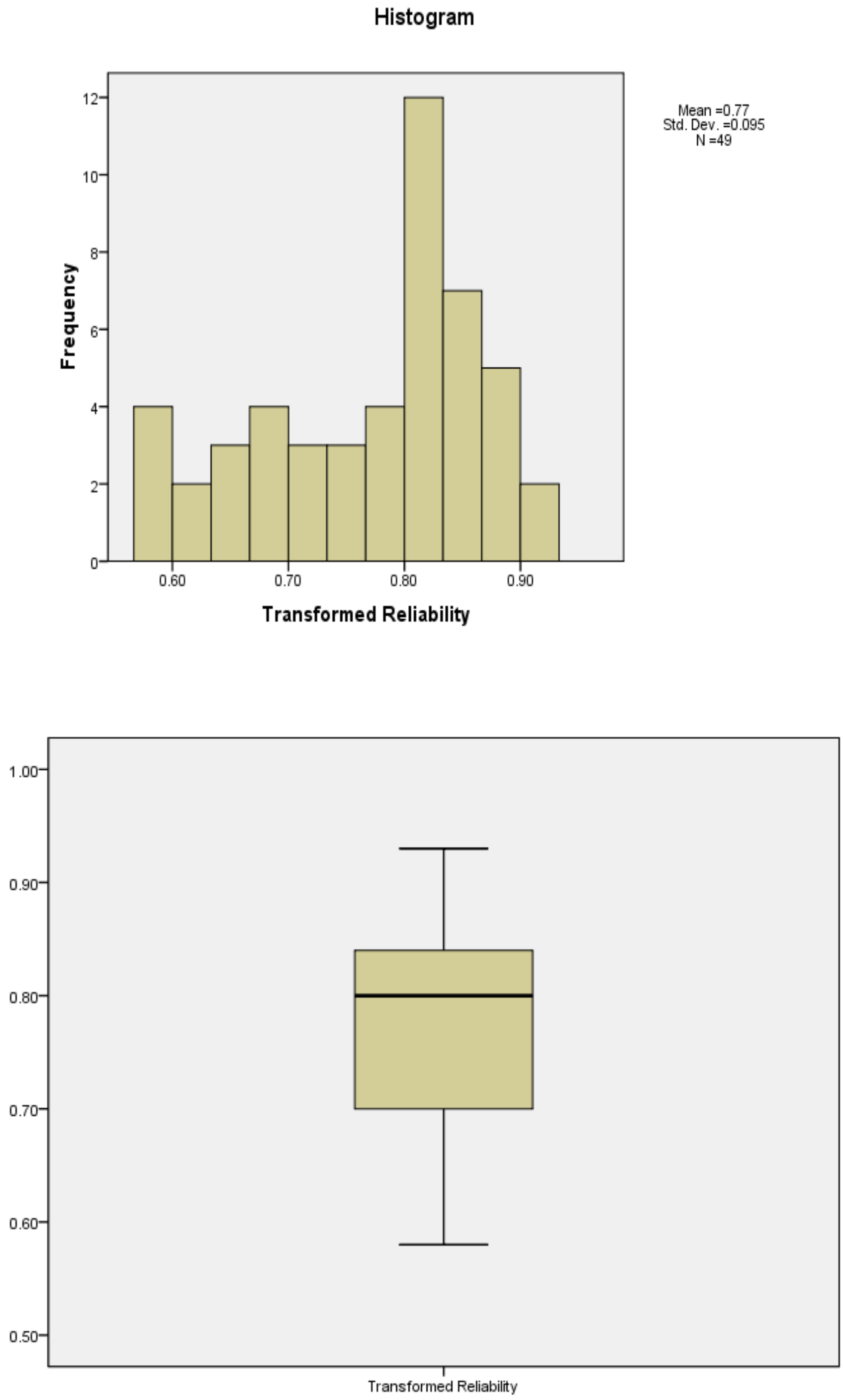
Histogram
Mean =0.77
Std. Dev. =0.095
N =49
Frequency
12
10
8
6
4
2
0
0.60
0.70
0.80
0.90
Transformed Reliability
1.00
0.90
0.80
0.70
0.60
0.50
Transformed Reliability

Figure 1.