Understanding and Using The Implicit Association Test: II. Methodological Issues

Brian A. Nosek

University of Virginia

Anthony G. Greenwald

University of Washington

Mahzarin R. Banaji

Harvard University

Authors' note: This research was supported by grants from the National Institute of Mental Health, MH-41328, MH-01533, and MH-57672 and the National Science Foundation, SBR-9422241 and SBR-9709924. The authors are grateful to Scott Akalis, Claudiu Dimofte, Jeff Ebert, Kristin Lane, Sriram Natarajan, and Eric Uhlmann for comments on an earlier version of this paper, and to Mary Hunt, Roger Despres, Joe Bella, Henry Kaufman, Philip Long, the Southern Poverty Law Center, and Yale University for their support of web-based research efforts. Correspondence concerning this article should be addressed to Brian Nosek, Department of Psychology, University of Virginia, Box 400400, Charlottesville, VA, USA, 22911. Electronic mail may be sent to nosek@virginia.edu.

Abstract

A measure of implicit social cognition, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), assesses relative strengths of associations between two target concepts (e.g., flowers and insects) and two poles of an attribute dimension (e.g., good and bad). Five studies used large datasets from demonstration Web sites to address methodological questions about the tool. Study 1 shows that data analytic strategies cannot relieve the procedural constraint of relative attitude measurement. Study 2 found that IAT effects were robust even with few stimuli (2 per category). Study 3 shows that the stimulus items chosen to represent target concepts can influence observed IAT effects. Study 4 suggests that the order of IAT and explicit (self-report) presentation does not influence either self-report or IAT effects or their correlation. Finally, Study 5 found that the known extraneous influence of order of IAT task blocks on IAT scores can be sharply reduced by introducing extra practice trials. Together, these experiments show some constraints in the IAT procedure that are difficult to remove and others that can be improved with simple modifications.

The resurgence of interest in the cognitive unconscious over the last 20 years may be attributable in part to the availability of a variety of new measurement tools. Measures of implicit cognition differ from self-report measures in that they can reveal mental associations that exist outside conscious awareness and are therefore inaccessible to explicit (or self-report) measures (Banaji, 2001; Bargh, 1997, 1999; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald & Banaji, 1995; Nosek, Banaji, & Greenwald, 2002c; Wilson, Lindsey, & Schooler, 2000). In recent years, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) has been used to study a diverse range of topics in part because of its ease of implementation, large effect sizes, and relatively good reliability (Greenwald & Nosek, 2001). Yet the ease of implementation of the IAT may incorrectly lead to the assumption that design, analysis, and interpretation of its effects are likewise simple. For example, Greenwald, Nosek, and Banaji (2003) presented an empirically derived scoring algorithm that substantially improves on the original scoring procedure that was introduced by Greenwald et al. (1998). The present paper continues that methodological effort by examining several questions concerning IAT procedures that have surfaced as experience with the tool has accumulated.

The popularity of the IAT as a research tool has produced a healthy degree of variation in its procedural and methodological features across studies and laboratories. However, significant degrees of procedural variation introduce several risks including: (a) methodological and procedural decisions may be seen as relatively unimportant and are determined without a strong rationale, (b) the effects of arbitrary procedural decisions on resulting data may go unnoticed, and (c) indiscriminant variation in procedural design may reduce the comparability of results across studies. By examining some routinely varying procedural elements, we identify some parameters for design of experiments using the IAT and describe the consequences of those methodological factors for the use of the IAT. Because some variation in procedural design can be healthy for the continuing development of a research tool, the paper's conclusions will not be a prescriptive guide for IAT design. Rather, this paper will provide an empirical rationale for some IAT design parameters that will enable researchers to make informed methodological decisions about their unique research objectives.

Implicit Association Test

In line with the basic tenets of theories of associative learning and representation, the IAT rests on the assumption that it ought to be easier to make the same behavioral response (a key press) to concepts that are strongly associated than to concepts that are weakly associated. The IAT procedure obliges respondents to identify stimulus items and categorize them into one of four superordinate categories. Strength of association is measured by comparing the speed of categorizing members of the superordinate categories under two different sorting conditions. For example, because the concepts 'Old' and 'Bad' tend to have stronger associative links than the concepts 'Old' and 'Good', respondents are able to identify and categorize items faster in a condition in which items representing 'Old' and 'Bad' share the same response compared to a condition in which items representing 'Old' and 'Good' share the same response.

The IAT's procedure has five steps, with steps 3 and 5 providing critical data.

Step 1: Learning the concept dimension. First, respondents sort items from two different concepts into their superordinate categories (e.g., faces of young people for 'Young' and faces of old people for 'Old'). Categorizations are made using two keys on a computer keyboard, which are mapped to the superordinate categories (e.g., the 'a' key for 'Old'; the ';' key for 'Young') and stimulus items appear sequentially in the middle of the computer screen.

Step 2: Learning the attribute dimension. In Step 2, respondents perform the same task with the same two keys, but now sort items representing two poles of an attribute dimension (e.g., terrible, nasty for 'Bad' and wonderful, beautiful for 'Good').

Step 3: Performing Pairing 1 (critical trials). In the third stage, these two sorting tasks are combined such that, on alternating trials, respondents are identifying a face as Old or Young, and then a word as Good or Bad. In this case, one key ('a') is the correct response for two categories (Old and Good) and the other key (';') is the correct response for the other two categories (Young and Bad). Respondents first perform a block of 20 trials with these sorting rules. After a brief pause, they perform the same sorting task for a second block of 40 trials.

Step 4: Learning to switch the spatial location of the concepts. In the fourth stage of the task, only stimulus items for the target concepts (Old and Young) are sorted for 20 trials, but this time the key assignment is reversed. In the present example, Old items would now require a ';' key response and Young items would require a 'a' key response.

Step 5: Performing Pairing 2 (critical trials). In the fifth stage of the task, respondents sort items from both the attribute and target concept categories again, except that the response key assignments now require that Young and Good items are categorized with one key, and Old and Bad items are categorized with the other key, the opposite association from the earlier block. As previously, respondents sort stimulus items with this response assignment for 20 trials, then have a short break, and then again for 40 more trials.

The IAT effect is calculated using latency data from Steps 3 and 5. In the above example, sorting the stimulus items faster when Old and Bad (and Young and Good) share a response key than the reverse pairings indicates a stronger association strength between Old and Bad (and Young and Good) compared to the reverse mapping, or an automatic preference for Young relative to Old. Greenwald et al. (2003) describe the scoring algorithm for calculating the IAT effect in detail. Essentially, the scoring algorithm involves calculating the difference in average response latency between the two sorting conditions and dividing by the standard deviation of all latencies for both sorting tasks. Thus, the IAT score (*D*) is a cousin of Cohen's d calculation of effect size for an individual's responses in the task.[1]

Goals

This article addresses the following five questions:

(1) **Can analytic methods separate the IAT's measure of relative association strength into two independent measures of association strength?** Any measure of attitude is obtained in a particular context. For example, when attitudes toward women are sought, the items often refer directly, indirectly, or merely imply a comparison to men. The IAT is conceptually similar to an explicit measure that specifically asks a question about say women's rights in comparison to men's rights. That is, it specifically creates a bi-polar dimension of concepts and obtains a relative assessment of the attitude toward the two, e.g., female as compared to male, which can be equally viewed as male compared to female.

This feature of the IAT has received much discussion regarding its implications for the interpretation of the score. One question in particular, concerns the possibility of seeking an assessment that is not a relative association between two target concepts, but the independent assessment of a single concept. For example, distinguishing in-group favoritism (e.g., Democrats liking Democrats) from out-group derogation (e.g., Democrats disliking Republicans) requires that evaluations of the in-group and out-group be assessed independently. Some researchers have tried analyses of subsets of IAT trials to obtain separate evaluations of two target concepts in the IAT, but it is unclear that the parameters of the tool allow any meaningful breakdown of two concepts as if they represented independent attitudes. Study 1 tests the appropriateness of this analytic strategy.

(2) **Is there an optimal number of stimulus items per category in the IAT?** Decisions about the number of stimuli used to represent each category in the IAT are often arbitrary, and at other times the number of available appropriate exemplars for a category dictates the size of the set. Study 2 examines the impact on IAT effects of varying the stimulus set size to represent categories.

(3) **Do the exemplars that represent a category influence IAT effects?** Since the inception of the IAT, a recurring question has concerned the degree to which the labels used in the classification task (e.g., Black-White, Old-Young) versus the particular exemplars of the category (e.g., words, pictures, faces), determine the IAT result. One prevailing interpretation is that IAT effects are determined more by the category identifications in the task than by the stimuli used to represent those categories (De Houwer, 2001; Nosek & Greenwald, 2001; Fazio & Olson, 2003). At the same time, several studies have indicated that semantic or evaluative properties of the items chosen to represent categories are important (Govan & Williams, 2003; Steffens & Plewe, 2001). Yet others have shown that both are crucial in producing the final result (Mitchell, Nosek, & Banaji, 2003). Study 3 investigates the influence of varying the features of exemplars used to represent categories on IAT effects.

(4) **Is the IAT result sensitive to the order in which it appears vis a vis self-report measures?** In studies that use both IAT and self-report measures, decisions must be made about the order of presenting study materials. For those primarily interested in IAT data, those responses are collected first, with explicit, self-report measures always appearing at the end. Others have collected self-report data first, based on the assumption that IAT data are less vulnerable to influence by self-report data rather than vice versa. Demonstrations of the malleability of implicit effects (e.g., Blair, Ma, & Lenton, 2001) suggest that such an assumption may not be valid. Study 4 tested the impact of order of measuring IAT and explicit measures on both types of measures.

(5) **Can the unwanted influence of order of IAT subtasks be avoided?** From the very first uses of the IAT, it has been known that although the IAT effect is robust, it is almost always larger when the dominant pairing is completed first. For example, in an age IAT, participants who first sort 'Old+Bad' and 'Young+Good' and then sort the reverse configuration show a stronger overall effect compared with participants who first sorted 'Old+Good' and 'Young+Bad'. More generally, the associations measured in the first of the two combined tasks tend to appear stronger than those measured in the second. This extraneous effect of order is observed in most IAT studies that vary this procedural factor (Greenwald & Nosek, 2001). Study 5 tests a procedural change designed to reduce this effect of pairing order.

Data source and data treatment

Two Web sites (http://www.yale.edu/implicit/ and http://tolerance.org/)[2] served as the data source for the five studies reported in this paper. The demonstration IAT portion of these sites functioned as the Internet equivalent of an interactive exhibit at a science museum. The Web sites provided visitors with a palpable experience of their own automatic associations that are unwanted, uncontrollable, or exist outside of conscious awareness. Completion of a test and a brief questionnaire required 5 to 10 minutes. Afterwards, respondents received a summary of their task performance along with a summary of results for large numbers of previous respondents. In addition to typical debriefing materials, the Web site provided information about implicit social cognition and provided answers to questions that had been repeatedly raised in email communications from prior visitors to the site. For further discussion about this use of Internet-based data see Greenwald et al., 2003 for its application to deriving a superior

scoring algorithm, and Kraut et al. (2003) and Nosek, Banaji, & Greenwald (2002a) for issues concerning the use of the Internet for data gathering more generally.

The large samples used for the present research (Ns range from 4,447 – 27,220; see Table 1) provide sufficient statistical power to guarantee that virtually all inferential statistical tests will yield statistically significant results. In fact, all 11 samples in this paper have a power value exceeding .99 to detect effects of d = .10 (n required = 3,675), r = .10 (n = 1,828), and q = .10 (n = 3,677), with a two-tailed test at p = .05 (Cohen, 1988). Consequently, the emphasis will be on presenting results in terms of effect size, rather than as meeting a significance level in the traditional sense.

Three types of effect sizes are reported in this paper. Cohen's d is used to standardize the magnitude of difference between means. The product-moment correlation coefficient, r, is an effect size measure reflecting the strength of covariation between two variables. And finally, to estimate the magnitude of difference between two correlations coefficients the measure q is used. Cohen (1988) established conventions for interpreting effects as small, moderate, or large for each of these effect size measures: values of .1, .3, and .5 are considered small, medium, and large effects for both r and q, whereas .2, .5, and .8 are considered small, medium, and large effects for d. As a guideline for this paper, r and q effects under .05, and d effects under .10 are not discussed in detail for their implications for measurement, analysis, and interpretation of IAT effects.

Method

Characteristics of respondents

Respondents found the publicly available Web sites through a variety of means including media coverage, hyperlinks from other sites, recommendations from friends or family, links from search engines, or word of mouth. Data from four tasks available at the Yale IAT Web site (formerly http://www.yale.edu/implicit/), and seven tasks available at http://tolerance.org, provided the samples for the five studies in this paper (see Table 1). All of the data were collected between 11/15/2001 and 10/29/2002. Approximately 90% of respondents answered at least one of the demographic items. Of those responding, 66% were female and 34% were male; 1% were American Indian, 5% were Asian, 6% were Black, 5% were Hispanic, 74% were White, 1% were Biracial (Black and White), 4% were multiracial, and 3% were other ethnicities; 60% were under 25 years of age, 35% were between 25 and 50, and 4% were over 50; 29% had a high school diploma or less, 57% had some college or a bachelor's degree, and 14% had an advanced degree; 89% were from the United States, 6% were from Britain, Canada, or Australia, and 5% were from other foreign countries; 48% considered themselves to be slightly to strongly liberal, 31% moderate, and 21% slightly to strongly conservative. Across the 11 tasks, 131,970 completed IATs were included in the data set.

Materials, apparatus, and evaluation criteria

IATs were presented via the Internet using applet and CGI technology. After a respondent elected to complete a task, a small program (Java applet) was automatically downloaded to the respondent's computer. The program used the respondent's computer resources to present stimuli and record response latencies, which avoided dependence on the speed of the Internet connection for accuracy of response latency measurement. Accuracy of response latencies with Java applets is limited to the operating system's clock rate (e.g., 18.2 Hz for Windows-based machines). This limitation was not an obstacle for implementing the IAT because of the nonsystematic nature of the resulting noise, the strong effects elicited by the IAT, and the substantial reduction of error magnitude achieved by averaging data over a substantial number of trials. Between blocks of task performance the applet returned data to the server database. Upon completion of the task, a summary score was calculated and reported to the

respondent as showing a 'strong', 'medium', 'slight', or 'little or no' strength of one of the association contrasts measured by the task. These labels correspond to results meeting the conventional criteria for small, medium, and large effect sizes of Cohen's d (1988) measure.

Implicit tasks. The data from the IATs used to answer the five questions posed for investigation were based on a variety of tasks that broadly measured attitudes and stereotypes toward social groups and individuals and covered race/ethnicity, age, gender, sexual orientation, and physical features. As a result, the conclusions we draw can be viewed as quite general. Each of the tasks used is described in brief, specifying the concept and attribute involved and the concept it is assumed to measure. All tasks are currently active at implicit.harvard.edu or www.tolerance.org and can be sampled by readers.

(1) Ethnic-national stereotype 1 (Native/White): the strength of association between Native American/White American faces and the attribute American/Foreign,

(2) Ethnic-national stereotype 2 (Asian/White): the strength of association between European/Asian faces and the attribute American/Foreign,

(3) Race attitude (Black/White): strength of association between African Americans/European Americans with the attribute Good/Bad,

(4) Skin color attitude (Dark/Light): strength of association between dark-skinned/light-skinned people with the attribute Good/Bad,

(5) Sexual orientation attitude (Gay-Straight): strength of association between gay/straight people with the attribute Good/Bad,

(6) Gender-Academic Domain stereotype (Male-Female): strength of association between Male/Female and the attribute Science/Liberal Arts,

(7) Age attitude (Young-Elderly): strength of association between Elderly/Young with the attribute Good/Bad

(8) Political candidate attitude (Gore-Bush): strength of association between Al Gore/George W. Bush and the attribute Good/Bad.

Three of the tasks (Old-Young, Gender-Science, Black-White) appeared at both demonstration Web sites. Summary data for each task are listed in Table 1. Calculation of the IAT *D* effect followed the procedures described by Greenwald et al. (2003), and contained the following main features: (1) error trials were removed and replaced with the mean of that performance block plus a penalty of 600ms, and (2) individual trial response latencies below 400ms were removed before analysis. Also, respondents for whom > 10% of the trial responses were less than 300ms were removed from the analysis (see the 'Excluded N' column in Table 1). The standard deviation of all response trials was used to calculate IAT *D* effects, even for analyses using only subsets of the task trials. This introduced the possibility of extreme D scores in cases of unstable responding between data subsets. So, an additional exclusion criterion of dropping IAT *D* effects with an absolute value of larger than 4.5 (at least 6 SD's or more away from the mean) was added to eliminate very extreme outliers. This correction resulted in the removal of approximately 0.4% of the data.

Self-report measures. Associated with each IAT, respondents received an optional questionnaire that included items assessing their attitudes or beliefs about the target categories, and some demographic information. Three items were of direct interest for the present studies. For each task, respondents rated their explicit preferences or stereotypes toward the target categories on a 5 or 9-point Likert scale. For example, for the attitude measure for Old-Young preferences, the Likert scale ranged from "I strongly prefer young people to old people" to "I

strongly prefer old people to young people." Stereotype measures were similarly phrased. For example, the Gender-Science Likert scale ranged from "I strongly associate females with liberal arts and males with science" to "I strongly associate males with liberal arts and females with science." The two other explicit items were 11-point thermometer scales in which the two target concepts were rated independently. For example, respondents answered the following for Old people and Young people: "Please rate how warm or cold you feel toward the following groups (0 = coldest feelings, 5 = neutral feelings, 10 = warmest feelings)." Likewise, for the stereotype items, respondents made two independent ratings on 11-point scales. For example, in the Gender-Science task respondents answered the following for Science and Liberal Arts: "Please rate how much you associate the following domains with Males and Females." Endpoints were 'Strongly Male' and 'Strongly Female'. A difference score between the two thermometer scales served as the thermometer score. The Likert score and the Thermometer difference score were standardized and then averaged to create a single explicit attitude or stereotype score.

Zero-order correlations between IAT effects and explicit measures served as a criterion for evaluating IAT measures. Previous research has documented a reliable, but variable relationship between implicit and explicit attitudes (Cunningham, Preacher, & Banaji, 2001; Greenwald & Farnham, 2000; Greenwald et al., 2003; Nosek, 2002; Nosek & Banaji, 2002; Nosek, Banaji, & Greenwald, 2002b; Nosek et al., 2002c) ranging from near zero to .78 (Bush-Gore attitudes; Greenwald et al., 2003). As such, assuming that the manipulated methodological variations in the implicit measures are not taking advantage of increased method covariance with explicit measures (unlikely given the significant procedural differences between implicit and explicit measures), the best measures of implicit and explicit attitudes will minimize random error and maximize relations with known covariates. For Studies 2 – 6, higher implicit-explicit correlations indicated more effective measurement of implicit attitudes as measured by the IAT (see Greenwald et al., 2003 for a detailed rationale for this approach).

Internal consistency. A measure of internal consistency was calculated by correlating IAT *D* effects calculated for the first block of 20 trials of the two pairing conditions (sometimes referred to as 'practice') and the second block of 40 trials of the two pairing conditions (sometimes referred to as 'test'). As an evaluation criterion, higher internal consistencies were considered desirable. In Study 5, internal consistency was calculated for self-reported attitudes and stereotypes by calculating the correlation between the Likert and Thermometer scores described in the previous section.

Extraneous influences. Resistance to effects of three known extraneous influences on IAT effects – average response latency, task pairing order, and experience with the IAT – served as evaluation criteria (see also Greenwald et al., 2003). Better performing IAT measures should be less affected by these extraneous influences. Average response latency was calculated by averaging response latencies across all trials in the two pairing conditions. Task pairing order was a between-subjects factor in which respondents were assigned to perform the two pairing conditions in one of two orders (e.g., Old+Bad/Young+Good first, or Old+Good/Young+Bad first). Also, respondents were asked if they had previously performed an IAT. The experience with the IAT factor was coded as a dichotomous variable of no previous experience versus some previous experience. Correlations (Pearson *r*'s) between the extraneous factors and IAT effects served as evaluation criteria with values closer to zero suggesting better performance.

Procedure

Each implicit measure was accompanied by a short questionnaire in which respondents reported explicit preferences for the target concepts and provided some simple demographic information. At the end of the session, respondents received a summary of their result, information about how others performed on the same task, and information about the research area of implicit social cognition and the task itself (see Nosek et al., 2002b and Greenwald et al.,

2003 for additional discussion of procedures for data collection, analysis, and handling of this type of data).

Limitations of web data

Self-selection. Because the data were collected at demonstration Web sites that were publicly available and primarily for educational purposes, the sample was self-selected. As a consequence, the sample cannot be said to be representative of any definable population. The absolute levels of implicit and explicit attitudes obtained therefore cannot be said to reflect the views of Americans as a group. Yet the very large N, greater diversity than most types of data collections (see respondent characteristics above), diversity of measures (8 unique, 11 total tasks), emphasis on hypothesis testing, and focus on methodological questions makes this dataset appropriate and useful for the current purposes for which methodological issues are the focus of attention.

Multiple participations by individual respondents. Visitors to the demonstration Web sites were encouraged to try as many of the tasks as they wished, and as many times as they wanted. Multiple data points from single respondents pose obvious threats for statistical analyses that make the assumption of independence of observations. However, the overall large number of respondents for each task reduces the potential impact of this threat. Also, data from these Web sites has been rigorously tested with regard to the influence of multiple data points from single respondents, with results indicating that overall effects are relatively unaffected by the inclusion or exclusion of multiple task performances from single respondents (see Nosek et al., 2002b for a more detailed discussion). One of the questions from the preliminary survey asked respondents to report how many IATs they had previously completed. This measure is used to test the effect of prior experience on IAT performance.

Study 1: Can analytic methods separate the IAT's measure of relative association strength into two independent measures of association strength?

The IAT is assumed to be a measure of the relative strength of association between concept-attribute pairs. For example, in a Flower–Insect attitude IAT, the measure yields the combined strength of Flower+Good/Insect+Bad associations and the strength of Flower+Bad/Insect+Good.

In the two critical conditions of the IAT, respondents sort items representing four concepts such as Flowers (e.g., daffodil, rose), Insects (e.g., fly, mosquito), Good (e.g., wonderful, beautiful), and Bad (e.g., terrible, awful) into two groups by categorizing each item with one of two keys on a computer keyboard. In this example, one condition would require that respondents sort Flowers with Good terms on one key (Flowers+Good) and Insects with Bad terms on the second key (Insects+Bad). The second condition would require that respondents sort Flowers with Bad terms on one key (Flowers+Bad) and Insects with Good terms on the second key (Insects+Good). The difference in average response latency for all the trials in the first condition compared to all of the trials in the second condition produces the usual IAT measure.

However, researchers might reasonably be interested in a simpler comparison, such as the relative strength of Flowers+Good and Flowers+Bad, obtained separately from that for Insect+Good and Insect+Bad. In at least two recent publications, researchers have created analytic strategies to interpret a portion of data collected within the IAT and interpreted the result as revealing such simpler association measures (de Jong, Pasman, Kindt, & van den Hout, 2001; Gemar, Segal, Sagrati, & Kennedy, 2001).

Gemar et al. (2001) examined implicit associations involving self-other (Me and Not Me) and valence. In the hope of examining evaluations of self separately from evaluations of other, Gemar et al. extracted two scores from each condition. In one condition, they calculated average

response latencies for Me+Good responses and Not Me+Bad responses separately; in the second condition, they calculated average response latencies for Me+Bad responses and Not Me+Good responses separately. This way, the authors interpreted the comparison of Me+Good and Me+Bad latencies as an independent measure of self-evaluation; and the comparison of not Me+Good and Not Me+Bad as an independent measure of other-evaluation. de Jong et al. (2001) used a similar analysis strategy.

The question that needs to be addressed is whether such a strategy is legitimate, i.e., whether a breakdown of the two parts of the IAT task can yield a meaningful indicator of independent attitude measures of the two concepts. So far, the developers of the IAT have assumed that such a strategy is inappropriate because the structural properties of the task suggest relative measurement. However, the issue is best resolved by evidence and Study 1 undertakes such an analysis.

Method

Materials

Data for four tasks at the demonstration Web site were used to test whether the trial-subset analytic method can extract independent assessments of association strength with the IAT (Bush-Gore attitude, Black-White attitude, Gender-Science stereotype, and Old-Young attitude; See Table 1).

Calculating measures of relative versus independent association strengths with the IAT

Three IAT scores were calculated for each task following the recommendations outlined by Greenwald et al. (2003).[3] Using the Election 2000 task as an example, separate scores for Bush and Gore were calculated. Calculation of the 'relative' IAT score followed the standard format. That is, all trials were retained for calculation of the IAT effect. Calculation of the 'independent' Bush IAT score was accomplished by first deleting trial latencies involving categorization of items referring to Al Gore, and evaluative terms sharing a response key with Gore (independent IAT [X only]). Likewise, all trial latencies involving categorization of items referring to George Bush and evaluative terms sharing a response key with Bush were deleted before calculating the 'Gore' IAT effect (independent IAT [Y only]).

Analysis Strategy

To test whether the IAT can be decomposed into independent assessments of association strengths, we compared the standard and separate scores for the implicit and explicit measures. For each task, independent explicit ratings were collected for each attitude object. For example, participants rated their liking for George Bush and Al Gore on separate scales. A difference score between the two ratings comprised the relative explicit preference for Bush versus Gore. If the IAT can be decomposed, then 'independent' IATs should show the strongest correspondence with independent explicit assessments of the same attitude object, whereas the 'relative' IAT should show the strongest correspondence with relative explicit assessments between the two attitude objects. That is, the independent 'Bush' IAT should show the strongest relationship with explicit Bush attitudes, and somewhat weaker relationships with explicit Gore attitudes and explicit Bush relative to Gore attitudes. Likewise, the independent 'Gore' IAT should show the strongest relationship with explicit Gore attitudes, and somewhat weaker relationships with explicit Bush attitudes and explicit Bush relative to Gore attitudes. Finally, the relative 'Bush versus Gore' IAT should show the strongest relationship with explicit Bush versus Gore attitudes, and somewhat weaker relationships with the independent explicit Bush attitude and explicit Gore attitude. This decomposable prediction is presented visually in the top right panel of Figure 1.

If the IAT is not decomposable, then parsing the data into a 'Bush' IAT and a 'Gore' IAT will be irrelevant for the nature of the associations measured – the 'independent' and 'relative' IAT measures will all be measuring relative attitudes. As a consequence, we would expect that all three IATs (Bush IAT, Gore IAT, Bush versus Gore IAT) would show the strongest relationship with the relative Bush versus Gore explicit attitude measure, and somewhat weaker relationships with explicit Bush attitudes and explicit Gore attitudes. This would occur because the subscores based on responses for a single attitude object [Bush] will still reflect relative associations between the two attitude objects [Bush versus Gore]. The pattern of correlations for the non-decomposable hypothesis is presented in the top left panel of Figure 1.

Results and Discussion

The four panels at the bottom of Figure 1 present the patterns of correlations for four tasks between the independent and relative explicit attitude measures and the 'independent' and 'relative' IATs for each task. If the IAT can be decomposed into independent attitude measures than correspondence should be maximized when IAT score and the explicit measure are matched (pattern in top right panel). If the IAT is not decomposable, then correspondence should not vary as a function of the IAT score (parallel lines as shown in top left panel). The pattern across tasks clearly matches the idealized non-decomposable pattern indicating that the correspondence between the IAT and criterion variables does not vary as a function of the match between measurement types as predicted by the decomposable hypothesis.[4] We conclude that *the IAT cannot be analytically decomposed into separate assessments of association strengths.*

While the data clearly indicate that the decomposing analytic strategy does not eliminate the relative assessment property of the IAT, they do not suggest that all IATs are therefore necessarily relative in their automatic attitude assessment. If both attitude objects equally influence IAT effects, it would be reasonable to predict that the correlations between the IAT and each attitude object in isolation would be of similar magnitude. Observation of the data patterns indicates that this is not always the case. For example, explicit Black attitudes are more strongly related to implicit Black-White attitudes (average $\underline{r}$ = .20) than are explicit White attitudes (average $\underline{r}$ = .10). Although certainly a function of both attitude objects, it is possible that Black-White IAT effects are proportionally more a function of attitudes toward Blacks than toward Whites. Strong conclusions about the nature of the IAT effect are not warranted by these data because the two explicit attitude reports could also be affected by numerous factors (e.g., differential response distortion, different variability in responses to one target concept compared to the other), which could lead to a differential relationship with the IAT effects. These data suggest that, if there is any control over the relative versus independent assessment with the IAT, it will be determined by design decisions not by analytic strategies.

Conclusion. Study 1 provides evidence that the block-decomposition strategies ventured by some researchers do not suffice to allow separate assessment of association strengths for the IAT's target concepts. This suggests that each trial response in the IAT, whatever the category membership of the exemplar, reflects some aspect of the relative comparison between the two target concepts. In addition, the data indicate that the predictive utility of the IAT will be maximized when criterion variables parallel the relativity built into the IAT. Researchers interested in using an implicit measure to examine non-relativistic preferences should consider a measure that is specifically designed for such purposes such at the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001).

Study 2: Is there an optimal number of stimulus items per category in the IAT?

Greenwald et al. (1998) noted that IAT effect magnitudes were unchanged when using only 5 exemplars per category compared to 25. Since then, little attention has been paid to the number of stimuli needed for designing an effective IAT. As a consequence, across tasks and

laboratories, the number of items representing each category varies from 2 to 30, with little basis for knowing whether this variation has any impact on observed effects. On the one hand, increasing the number of exemplars may assist in providing a more accurate representation of its superordinate category, and reduce the likelihood that respondents could learn to categorize stimuli on the basis of feature recognition rather than the meaning of the concept. On the other hand, decreasing the number of exemplars may help to avoid diluting the representation of the superordinate category by avoiding stimulus items that do not directly capture the category meaning.

In Study 2, using data from three of the tasks described in Study 1, we varied the number of exemplars representing the categories comprising the task and measured its effect on the magnitude of the IAT effect, reliability, relationship with self-reported attitudes, and resistance to extraneous influences (i.e., average response latency, order of task pairings, IAT experience).

Method

Materials

Data from three of the tasks in Study 1 (Black-White, Gender-Science, Old-Young from the Demonstration Web site) comprised the data for Study 2. The Black-White task consisted of 6 exemplars (i.e., black and white faces) for the Black and White categories and 8 exemplars (e.g., wonderful, terrible) for the Good and Bad attributes for a total of 28 items. The Gender-Science task consisted of 8 exemplars for each of the four categories: Male, Female, Science, and Liberal Arts (e.g., man, woman, physics, history) for a total of 32 items. The Old-Young task consisted of 6 exemplars (i.e., old and young faces) for the Old and Young categories, and 8 exemplars for the Good and Bad attributes for a total of 28 items.

Procedure

After selecting a task to perform, respondents were randomly assigned to a between-subjects condition that varied the number of stimulus items. For the Black-White task, the 6 exemplars representing the Black and White categories was held constant across respondents, and the number of items representing the Good and Bad attribute categories was manipulated across four conditions: 8 exemplars each (28 total stimuli), 4 exemplars (20 total), 2 exemplars (16 total), and 1 exemplar [category label 'Good' and 'Bad' only] (14 total).[5] For the Gender-Science task, the number of stimuli for both categories and attributes was manipulated simultaneously across four conditions: 8 exemplars each (32 total stimuli), 4 exemplars each (16 total), 2 exemplars each (8 total), and 1 exemplar [category labels 'Male', 'Female', 'Science', 'Liberal Arts' only] (4 total). For the Old-Young task, the 8 items representing the Good and Bad attribute categories were held constant across conditions, and the number of items representing 'Old' and 'Young' was manipulated across two conditions: 6 exemplars each (28 total stimuli), and 2 exemplars each (20 total stimuli). For conditions in which only a subset of the total stimuli were used (e.g., 4 of the 8 available exemplars), additional between subjects conditions were introduced to ensure that all exemplars were used in equivalent amounts for each of the conditions. This way, differences between conditions cannot be due to the particular items selected to represent the categories.

Results and Discussion

Table 2 presents effect magnitude, relation to self-report, reliability, and resistance to extraneous influence data for the three tasks separated by the 'number of exemplars' conditions. Somewhat surprisingly, the number of exemplars used to represent the categories did not have a strong impact on IAT effects until very few stimuli were used per condition. Effect sizes for the Black-White task showed only slight variation whether 8, 4, 2, or 1 exemplar(s) represented the categories 'Good' and 'Bad'. Also, the effect magnitude for the Old-Young task declined only slightly when the 6 exemplars representing each of the categories 'Old' and 'Young' were

reduced to only 2 per category (d = .09). Effects did begin to decline more noticeably in the Gender-Science task when just two stimuli were used to represent each of the four categories (d = .85 compared to d = .97 when four stimuli represented each category), and then showed additional declines in the 1-exemplar condition in which only the category labels served as exemplars (d = .60). These data suggest that, until only a minimal number of stimuli are used per category (2 or 1), IAT effect magnitudes are relatively unaffected by the number of exemplars representing each category.

Across tasks, relationships between the IAT and self-reported attitudes were relatively consistent whether 8, 4, or 2 items represented the categories. The average variation in implicit-explicit correlations of q = .04 across tasks is almost completely explained by lower correlations for the 1-exemplar condition, when only the category label was used. Likewise, split-half reliabilities and relationships between attribute and category effects were very stable with the exception of the 1-exemplar conditions, which were noticeably lower.

Varying the number of stimuli representing each category had little effect on three known extraneous influences of IAT effects. While effects of pairing order varied somewhat depending on how many stimuli represented, those effects were not consistent across tasks: the effects of pairing order appeared to increase with fewer stimuli for the Old-Young task (r = .07 to r = .19), but decrease with fewer stimuli for the Gender-Science task (average r = .16 for '8', '4', and '2' conditions to r = .03 for '1' condition) and the Black-White task (average r = .09 for '8', '4', and '2' conditions to r = .04 for '1' condition).

Conclusion. Variation in the number of exemplars representing the attributes and categories had little impact on effect magnitude, reliability, or relations with self-report until categories were represented by only a single exemplar – the category label. Even in this circumstance, interpretable, though weaker, IAT effects did emerge. IATs with only 2 items representing each attribute and category showed effects only slightly less robust those using 8 items per attribute and category, suggesting that design decisions concerning the number of items for each category should focus more on selecting items that best represent the category, and less on generating large sets of stimulus items. In Study 4, we follow-up this observation with a more focused investigation of the importance of selecting appropriate exemplars to represent the attributes and categories of direct interest.

Study 3: Do the exemplars that represent a category influence IAT effects?

To date, our position regarding the selection of concept and attribute labels and stimuli necessary for the IAT has been that the labels should represent the attitude object of interest simply and clearly, and that the stimulus items should be selected to clearly belong to one, and only one, of the superordinate categories. When stimulus items are easily categorizable as more than one of the superordinate categories, the task demands of identification and categorization can interfere with the assessment of association between concepts and threaten the interpretability of the resulting effects.

However, more complex issues regarding these two dimensions, labels and stimuli, have been raised. The IAT requires the identification of individual exemplars of concept categories as belonging to one of two categories and variation is possible at both levels: the manner in which the concept/attribute category itself is specified (e. g., Black vs. White or African American vs. European American), and the types of exemplars used to represent the category (e.g., famous and liked individuals from the two groups, or unknown individuals). Do these design decisions have implications for the resulting task? For example, Mitchell et al. (2003) measured implicit attitudes toward Black athletes compared to White politicians with two separate IATs. One emphasized the race of the targets by using 'Black' and 'White' as category labels. The second used the same stimulus items, but emphasized the occupation of the targets by using 'Athlete'

and 'Politician' as category labels. Despite the fact that the stimulus items were identical, the resulting implicit attitude varied as a function of the labeling of the categories. When race was emphasized, White politicians were preferred over the Black athletes, but when occupation was emphasized, Black athletes were preferred over the White politicians. These results suggest that category labels set the interpretive frame for stimulus items and that it's choice can determine the attitude effect that is observed. Category labels should represent the concepts of interest, be easily understood by the respondents, and provide clear distinctions between the four categories comprising an IAT.

While the specification of category labels in determining the IAT effects now seems obvious, the influence of the particular exemplars representing the concept categories is less understood and the goal of Study 3 is to address this issue more substantively. Mitchell et al. (2003) showed that the IAT race effect, when using the labels Black and White but with liked Black exemplars and disliked White exemplars produced a much weaker race bias ($\underline{d}$ = .23) than when the Black exemplars were disliked and the White exemplars were liked ($\underline{d}$ = 1.68). Yet, some researchers have argued that IAT effects are a function of the category labels, and not the individual exemplars (De Houwer, 2001; Fazio & Olson, 2003). Others have shown that exemplars do affect implicit responses when the meanings of stimulus items are confounded so that they represent more than one of the superordinate categories (Steffens & Plewe, 1999), and the question remains an open one with the data suggesting that both can play a role in the resulting IAT effect. The current study examines this question substantively with three types of stimulus variations to test whether selection of stimuli has important consequences for the magnitude, reliability, predictive utility, and resistance to extraneous influences of IAT effects.

It is clear from previous research that dramatic changes in stimulus qualities will influence IAT scores (Mitchell et al., 2003; Steffens & Plewe, 2001), so this study examines some conditions in which more subtle variations could have an effect. For three tasks, using subsets of a homogeneous set of items produced a subtle manipulation of stimulus features (e.g., different old and young faces; Black-White, Old-Young, and Gender-Science tasks). This manipulation should not elicit different IAT effects unless the task is extremely sensitive to irrelevant stimulus variations. A more significant variation of stimuli involved generating two unique sets of stimuli that were conceptually similar but physically distinct (pictures versus words; Native-White stereotype). IAT effects might vary with this task if stimulus modality is an important factor. And finally, a third test involved switching of just two exemplars from the stimulus set with slightly different conceptual meanings (Gay-Straight attitude). This task might show variation in IAT effects if subtle shifts in semantic meaning of exemplars can influence IAT effects.

Method

Materials

Implicit measures. For three of the tasks (Black-White, Gender-Science, and Old-Young at the Demonstration Web site) multiple versions of the tasks were created with subsets of exemplars taken from a single set of items. The two versions of the Black-White task used the same Black and White faces, but each used four of the eight good and bad terms from the original race IAT for the categories 'Good' and 'Bad.' The eight text stimuli for each of the four categories in the Gender-Science task were split in half to create two versions of the task with no stimuli in common.[6] Finally, the six faces representing the categories 'Old' and 'Young' in the Old-Young task were randomly divided into three sets of two to create three versions of that task. The eight stimuli representing the categories 'Good' and 'Bad' in the Old-Young task were held constant across versions of the task.

For the Native-American task, two tasks were created that were designed to represent the same concepts using very different stimulus modalities (pictures versus words). Both versions used 'historical' White and Native American faces to represent the categories 'White American' and 'Native American', but used different sets of stimuli to represent the categories 'American' and 'Foreign'. In one case, those categories were represented by names of American or foreign places (e.g., Utah, Miami, Oslo, Italy), and in the other case the categories were represented by pictures of natural scenes of American or foreign landmarks (e.g., Grand Canyon, Sahara Desert).[7]

The two versions of the Gay-Straight task subtly varied the emphasis of the category 'Gay' to refer to lesbians or gay men. In both tasks, the same eight stimuli were used to represent the categories 'Good' and 'Bad', the same four stimuli represented the category 'Straight', and two of the four stimuli for the category 'Gay' were held constant across tasks (i.e., the words 'Gay' and 'Homosexual'). The only difference between the two versions of the Gay-Straight task involved two exemplars representing the category 'Gay'. In one case, the two images represented lesbians (i.e., two wedding cake bride figurines together, and two female 'stick figures' side-by-side), and in the other case, the two images represented gay men (i.e., two wedding cake groom figurines together, and two male 'stick figures' side-by-side). In sum, the tasks were nearly identical except for the slight shift in emphasis of gender for two items representing the category 'Gay'.

For all tasks, task-version conditions were manipulated between subjects, and respondents were randomly assigned to condition.

Results and Discussion

Table 3 presents effect magnitudes, implicit-explicit correlations, internal consistencies, correlations between attributes and concepts, and correlations with known extraneous influences on five tasks each with 2 or 3 conditions in which different sets of exemplars represented the concepts of interest. For three tasks (Black-White, Gender-Science, Old-Young), the different versions of the tasks were created by using subsets of stimuli from the same stimulus set, which should maximize the task similarity except for the specific items used to represent the tasks. For the Native-American task, the two versions of the task were designed to capture the same concepts but varied in stimulus modality (picture or word). And, for the Gay-Straight task, the two versions of the task were nearly identical except that two stimuli in the category 'Gay' shifted the salience of lesbians or gay men. As can be observed in Table 4, across tasks, variations in stimuli had very little impact on most of the evaluation criteria: implicit-explicit correlations (average $\underline{q}$ = .02), internal consistency (average $\underline{q}$ = .05), attribute-concept correlations (average $\underline{q}$ = .07), and resistance to extraneous influences (overall speed average $\underline{q}$ = .07, pairing order average $\underline{q}$ = .06, experience with the IAT average $\underline{q}$ = .02).

From the above effects, we might be tempted to conclude that exemplars play little to no role in the production of IAT effects. And it is clear that modality differences (pictures versus words) or different samples of exemplars selected from the same population produce similar IAT effects. This result is reassuring because it suggests that the IAT effect is not sensitive to the particular items that are chosen from a larger set that is homogeneous. It also suggests that IAT effects across different studies and from different laboratories using comparable stimuli from a superordinate set can be meaningfully compared.

However, when the exemplars vary in the content such that their meaning evokes a different evaluation, the stimuli can be said to be measuring different attitudes and the resulting IAT effect should differ. For three of the tasks, variations in stimuli produced a very small effect on the magnitude of IAT effects (Black-White $\underline{d}$ = .03, Old-Young $\underline{d}$ = .09, Native-American $\underline{d}$ = .09). But, for the Gender-Science ($\underline{d}$ = .19) and especially for the Gay-Straight task ($\underline{d}$ = .28),

changes in the exemplars representing the categories led to noticeable differences in the magnitudes of the resulting effects (see Table 3). This is perhaps most surprising for the Gay-Straight task because only two of the total of 24 stimuli were different between the two tasks. Even so, negativity toward 'Gay' relative to 'Straight' was much stronger when the category 'Gay' made salient gay men rather than lesbians, fitting with much existing theory that anti-gay attitudes are stronger for gay men than lesbians. While it is notable that the implicit-explicit attitude relationship did not also change as a function of the gender salience manipulation, that effect may be a function of the particular construct of interest (i.e., Gay-Straight attitudes) and not a suggestion that similar effects will be observed for other attitude objects. While attitudes toward lesbians might be less negative than attitudes toward gay men, individual variation in liking might be very consistent. The effects of exemplars on correlations and prediction need to be explored further.

The results of this study clearly show two results of interest. First, trivial stimulus shifts from homogeneous sets of stimuli do not seem to influence IAT effects. Small variations in IAT scores were observed across subsets of good and bad exemplars (Black-White task), faces (Old-Young task), male and female pronouns, and science and humanities disciplines (Gender-Science task). The fact that stimulus modality (picture or word) did not have an impact on IAT effects is noteworthy. However, previous research has indicated an influence of stimulus modality suggesting that this effect should be examined further (Nosek et al., 2002b). Second, when the exemplars tap a subcategory of the category label that is evaluatively distinct from other subcategories (Gay-Straight task), the IAT score may reflect that shift, as it well should. Together these results suggest that stimulus items may influence IAT scores when the variations in stimulus features are evaluatively relevant, but may have little effect when the features are evaluatively irrelevant.

Conclusion. Selection of exemplars for categories in the IAT should be made carefully with an emphasis on selecting stimuli that are (a) easily identified as members of their superordinate category, (b) not members of multiple superordinate categories, (c) reinforce the concept or attitude object of interest, and (d) as a set, fully represent all aspects of the concept or attitude object of interest, and not inadvertently emphasize only a part of the concept (e.g., only lesbians if the more general category 'Gay People' is of interest). These data also suggest that very minor stimulus variations (Black-White, Old-Young, and Gender-Science tasks) and use of pictures versus words (Native-American task) do not necessarily alter IAT effects.[8]

Study 4: Is the IAT result sensitive to the order in which it appears vis a vis self-report measures?

The preceding studies focused on features of the IAT procedure itself. However, IATs are frequently administered alongside other measures such as questionnaires. Does increasing the accessibility of explicit attitudes or beliefs by administering explicit measures before implicit measures alter the effects observed on the IAT? Alternatively, does completing the IAT and perhaps gaining the insight that one is showing a bias in a particular direction influence responding on a subsequent explicit measure of preference or stereotype? Early studies with the IAT tended to fix the order of implicit and explicit measures with the explicit measures occurring first, the assumption being that performance on the IAT would be unaffected by completing explicit measures, but that the very nature of explicit measures would not warrant the same assumption. However, recent demonstrations of the malleability of implicit attitudes and beliefs (Blair et al., 2001; Dasgupta & Greenwald, 2001; Lowrey, Hardin, & Sinclair, 2001; Mitchell et al., 2003) suggest that implicit associations are influenced by previously performed tasks, contextual cues, and the immediate situation. Initial self-report of preferences may influence subsequent implicit assessment either through changing the context of measurement,

or by making explicit preferences temporarily accessible. In the current study, we directly test the influence of task order.

Method

The tasks from Study 2 (Black-White, Old-Young, and Gender-Science preferences from the Demonstration Web site) were used to test the effects of order of the task. After respondents selected a task, they were randomly assigned, in a between-subjects design, to receive the implicit or the explicit measure first.

Results and Discussion

Table 4 presents implicit and explicit effect magnitudes, relations between implicit and explicit measures, implicit and explicit internal consistencies, and relations between the IAT and extraneous factors for three tasks (Black-White, Gender-Science, Old-Young) separated by the order of presentation of IAT and explicit measures. Examination of the average effects across evaluation criteria reveals that the order of presentation of implicit and explicit materials had a minimal effect on implicit effects and explicit reports. In fact, average effects were near zero for all of the evaluation criteria including IAT effect magnitudes ($\underline{d}$ = .02), explicit effect magnitudes ($\underline{d}$ = .06), implicit-explicit correlations ($\underline{q}$ = .00), implicit internal consistency ($\underline{q}$ = .00), explicit internal consistency ($\underline{q}$ = .00), and resistance to extraneous effects of overall speed ($\underline{q}$ = .00), task pairing order ($\underline{q}$ = .02), and experience with the IAT ($\underline{q}$ = .02). While largely showing no effects due to order of implicit and explicit measures on average, small influences were observed on the magnitude of implicit ($\underline{d}$ = -.12) and explicit ($\underline{d}$ = .14) measures of Gender-Science stereotypes. The conclusion offered by the current evidence is that the order of implicit and explicit measures does not have a strong influence on any of the evaluation criteria for implicit or explicit measures.

Despite the consistent lack of influence of task order on implicit and explicit measures, interpretation of the effects observed here should be tempered by the specific nature of the measurement context. For one, before completing any of the measures, respondents were aware that stereotypes and attitudes toward a particular category were going to be measured. It is possible that stronger effects of task order would be observed if the target concepts were not made known until measurement commenced. Also, the implicit and explicit measures (including demographics questions) were relatively simple and straightforward. In all cases, the explicit measures consisted of fewer than 10 items and fewer than 8 items measuring demographic characteristics (e.g., age, sex, ethnicity). More involved or extended assessment of explicit preferences or beliefs may have a stronger effect of subsequently measured implicit attitudes by making those explicit preferences more accessible, or by some other psychological mechanism. For example, Blair et al. (2001) showed a reduction of implicit associations between the concepts female and weak when respondents first performed a writing exercise about strong women. If similar writing exercises were part of a battery of questionnaires, rather than an experimental manipulation as it was for Blair and colleagues, then the order of presentation of materials could have a strong influence on the degree of implicit stereotyping revealed by the IAT (see also Bosson, Swann, & Pennebaker, 2000). Finally, the three attitudes and stereotypes examined in this study are presumably well learned. The order of implicit and explicit measures may have a stronger effect on attitudes and stereotypes that are relatively new, unstable, or ambivalent because the self-report could make certain evaluations more accessible when explicit measures are first, or the experience of performing the implicit measures may be seen a relevant information for generating a self-reported preference when implicit measures are first.

Conclusion. The present study suggests that in relatively straightforward measurement contexts with well-learned attitudes, the order of presenting implicit and explicit measures has

little influence on the effects elicited by these two measurement types. In similar measurement contexts, researchers may make decisions about task order based on pragmatic rather than methodological considerations. However, until more extensive research is done, researchers should carefully consider task order for situations in which explicit measures are extensive or the attitude objects are relatively novel, unstable, or likely to elicit ambivalent responses.

Study 5: Can the unwanted influence of order of IAT subtasks be avoided?

A persistent methodological artifact for the IAT is the order of task performance blocks. For example, respondents who complete the condition where Insects are sorted with Bad and Flowers are sorted with Good before sorting Insects+Good and Flowers+Bad will show some interference in performing the second condition. As a consequence, those respondents will show a stronger preference for flowers relative to insects than respondents who first sort Insects+Good and Flowers+Bad. In this section, we test a methodological innovation to remove the effect of pairing order from the IAT.

To review, the standard format of the IAT involves five blocks of sorting trials. Following the current example, the first block would typically contain 20 trials of sorting Flowers and Insects into their respective categories. The second block contains 20 trials of sorting Good and Bad terms into their respective categories. The third block contains 20 trials of sorting Flowers+Good terms on one key and Insects+Bad terms on the second key, a pause, and then another 40 trials of the same sorting conditions. In the fourth block, respondents again sort Flowers and Insects for 20 trials, but this time on the opposite keys from the first three blocks. The fifth block contains 20 trials of sorting Insects+Good terms on one key and Flowers+Bad terms on the second key, a pause, and 40 more trials of the same sorting conditions. The third and fifth blocks described above are typically counterbalanced between subjects (giving rise to the task pairing order effect).

If the pairing order effect is due to the interference caused by learning an initial response set and then needing to replace it with a new response set, then practice with the new response set may reduce the effect of task order. Our strategy to eliminate the order effect was to increase the number of trials in the fourth block in which respondents are becoming accustomed to sorting the target concepts on the opposite keys from which they had been sorting those concepts in the previous blocks. In the following study, we varied the number of trials in the fourth block by increments of 5 trials between 20 and 40 trials for five different tasks. That is, different groups of subjects received 20, 25, 30, 35, or 40 practice trials in the fourth block of the IAT as described above. We tested whether adding trials to this block was sufficient to make the magnitude of the IAT effect equivalent regardless of the order in which the critical blocks were completed.

Method

Materials

Five tasks (Old-Young attitude, Asian-White stereotype, Gender-Science stereotype, Black-White attitude, and Dark-Light attitude) that were available at http://tolerance.org and summarized in Table 1 were used for this experiment.

Procedure

We varied the number of practice trials for the fourth block across five between-subjects conditions (20, 25, 30, 35, 40 trials). Each of the five conditions was administered multiple times in 1-3 week intervals for each of the five tasks. Also, at any given point in time, no two tasks were in the same practice trial condition.

Results and Discussion

The impact of pairing order on effect magnitude, reliability, relationship with explicit attitudes and stereotypes, and resistance to extraneous influences. Table 5 summarizes the evaluation criterion for five tasks with data separated by the pairing order of the IAT. The 'compatible' block was operationally defined as the response pairing that was completed fastest for the majority of respondents. The influence of pairing order on effect magnitude in this table is likely somewhat reduced as compared to previous analyses (column 'pairing order' r-values in Tables 2-5) because these data collapse across the conditions designed to reduce the impact of the order effect. Even so, larger effects tended to be observed when the 'compatible' block of the task was first compared to second (average d = .18). However, the influence of pairing order on other evaluation criterion was small and inconsistent. Performing the compatible task first was associated with slightly lower relations between implicit and explicit preferences (average q = .02), non-differentiated internal consistency (average q = .00), and only minimally reduced vulnerability to extraneous influences of overall speed (average q = -.01) and experience with the IAT (average q = -.03). The impact of pairing order seems to have its strongest influence on the magnitude of IAT effects and not on its reliability, relations to criterion, and vulnerability to some extraneous influences. Next, we examined whether increasing the number of reverse single-discrimination practice trials was sufficient to reduce or eliminate the effects of pairing order on effect magnitude.

Removing the order effect by adding trials to the single discrimination practice block in the second half of the IAT. Figure 2 presents the magnitude of the order effect by the number of trials in the reverse single discrimination practice block (20, 25, 30, 35, or 40 trials) for five tasks and the unweighted average of those tasks. Elimination of the order effect would be seen if the magnitude of the order effect was equal to 0. The sharp decline in effect magnitude across conditions shows that increasing the number of practice trials had a strong influence in reducing the order effect from an average r of .15 with 20 practice trials, to an average r of .03 with 40 practice trials. This simple change in the task procedure was sufficient to virtually eliminate the extraneous effects of pairing order.

Interestingly, while the order effect was certainly reduced for the Gender-Science task (r = .25 down to r = .17) it was not completely eliminated. There are too many differences between tasks to isolate the specific reason for the persistence of an order effect for the Gender-Science task. One obvious candidate, however, is that while most tasks used pictures or faces to represent the target categories and words to represent the attributes, the Gender-Science task was the only one that used words to represent all four categories. It is possible that tasks with only a single stimulus modality will show more unrelenting influence of pairing order than tasks with multiple stimulus modalities. However, there are other possible reasons for the persistent effect of pairing order making this hypothesis tentative rendering this comment a speculation.

Conclusion. Results from Study 5 suggest that adding additional practice to the reversed single discrimination practice block will reduce pairing order effects, and even eliminate them in many cases. The recommendation is to use such additional practice trials. Even so, it is imperative that order of the pairing tasks should continue to be counterbalanced to ensure that this extraneous factor will not exaggerate or underestimate effect magnitudes and subsequent interpretations.

General discussion

Five studies investigated methodological issues relevant to the design, analysis, and interpretation of the Implicit Association Test (IAT). The results add to existing basis for informed decisions about procedural design of studies using the IAT. The findings can be summarized as answers to the five questions presented at the beginning of the article:

(1) **Can analytic methods separate the IAT's measure of relative association strength into two independent measures of association strength?** Answer: No. Study 1 demonstrated that the relative nature of the IATs procedural format cannot be undone via analytic methods. Even when subsets of its trials are the focus of analysis, the IAT remains a relative measure of association strengths. This result reinforces the importance of selecting the appropriate comparison category in the IAT. Selection of a random, 'neutral', or irrelevant comparison may not fulfill researchers' expectations for eliciting an independent association for their target concept of interest. Researchers interested in assessing associations with a single target concept should use a method designed for that purpose (e.g., Nosek & Banaji, 2001) or an adaptation of the IAT designed to focus on associations with one category (e.g., Swanson, Rudman, & Greenwald, 2001).

(2) **Is there an optimal number of stimulus items per category in the IAT?** Answer: At least 2 per category, with qualifications. Study 2 showed that IAT effects could be observed with stimulus sets that are comprised only of the category labels for the task. This observation, however, comes with an important caveat that these IATs show less robust effects than tasks with at least 2 stimuli per category. Decisions about the number of stimuli to use for an IAT can be based on pragmatic concerns, with at least 4 stimulus items per category being ideal, but 2 items per category being sufficient. The most effective IATs will use stimulus items that are easily identified as members of the superordinate category, are not confounded with other categories in the task, and are representative of the concept of interest. Likewise, category labels should directly reflect the construct of interest and maximize the ease with which respondents can identify the category membership of each stimulus item.

(3) **Do the exemplars that represent a category influence IAT effects?** Answer: Yes. Selection of a stimulus set is not arbitrary and unimportant. Reassuringly, Study 3 demonstrated that subsets of items from the same stimulus set (e.g., pictures of non-smiling, old and young faces) elicit consistent and stable effects. Also, at least one variation across stimulus modalities (picture or word; Native-White stereotype) did not elicit dramatically different IAT effects, suggesting that the IAT effectively taps the semantic constructs of interest and may not be inordinately influenced by orthographic features. Study 3 also showed that subtle changes to the semantic content of the stimulus sets can have a significant impact on IAT effects. Changing just two of the 24 exemplars in a Gay-Straight attitude task from ones that depicted gay men to ones that depicted lesbians significantly affected the magnitude of the IAT effect. This result suggests that the role of stimulus items for IAT effects may be to provide contextual meaning for the superordinate categories. In the current example, the change of exemplars may have shifted the meaning of the category 'Gay' across gender lines.

(4) **Is the IAT result sensitive to the order in which it appears vis a vis self-report measures?** Answer: No, with important qualifications. In Study 4, little to no differences were observed in implicit and explicit effects between presentation orders. This observation suggests that task order is unimportant when the study materials are simple and the target concepts are well defined and well known. This conclusion, however, is constrained by the narrow range of circumstances tested. It is clear that situational or contextual factors influence performance on implicit measures (Blair, 2002) making it unwise to conclude that task order is irrelevant for implicit measurement. Concern about the presentation order of implicit and explicit measures may be unnecessary when the implicit and explicit measures are relatively short and simple, and where responses to the target concepts are likely to be stable and unambivalent. For other situations, careful attention should be paid to presentation order of implicit and explicit measures.

(5) **Can the unwanted influence of order of IAT subtasks be avoided?** Answer: Yes, with qualifications. One of the most robust and well-documented extraneous influences on the IAT is the order of task performance blocks (task pairing order). In Study 5, we demonstrated this effect of pairing order and offered a procedural change to dramatically reduce its influence. Doubling the number of trials in the reverse single-discrimination block of trials (Block 4) from 20 to 40 reduced the overall impact of task order to $\underline{r} = .03$. This procedural change has the practical significance of minimizing the influence of a significant extraneous factor on IAT effects. Importantly, for one task, Gender-Science stereotype, the order effect did decline with this procedural adjustment, but it did not disappear. We hypothesized that pairing order effects may be more robust with IATs using lexical stimuli exclusively. This conclusion is necessarily speculative until follow-up research is conducted. Pragmatically speaking, this observation suggests that researchers will benefit from introducing this procedural variation in their own designs, but that they should continue to counterbalance the order of task performance blocks in case lingering effects of order occur.

Conclusion

The inextricable link between theory and method in science makes the rigorous examination of method of critical importance for the advancement of theory. Psychological theories, without appropriate empirical methods to test them, have nowhere to go but the humanities. Pragmatically, attention to methodological questions will minimize the amount of effort wasted on research projects yielding weak, uninterpretable, or inexplicably contradictory results. We may then be left with the methodologically sound but ego-depleting scenario that research failures can be blamed only on poor theorizing. In the present paper, five studies presented data with pragmatic implications for the design, analysis, and interpretation of the Implicit Association Test. With much still to learn about the IAT, we hope that this paper will accelerate the process of methodological and theoretical discovery in implicit social cognition.

References

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

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

Bargh, J.A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), Dual-Process Theories in Social Psychology (pp. 361-382). New York, NY: Guilford Press.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. Personality and Social Psychology Review, 6, 242-261.

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

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

Cohen, J. (1988). Statistical power analysis for the Behavioral Sciences. Lawrence Erlbaum Associates: Hillsdale, NJ.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. Psychological Science, 12, 163-170.

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

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. Journal of Experimental Social Psychology, 37, 443-451.

de Jong, P. J, Pasman, W, Kindt, M, & van den Hout, M. A. (2001). A reaction time paradigm to assess (implicit) complaint-specific dysfunctional beliefs. Behaviour Research & Therapy, 39(1), 101-113.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? Journal of Personality and Social Psychology, 69(6), 1013-1027.

Fazio, R. H. & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. Annual Review of Psychology, 54, 297-327.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. Journal of Personality and Social Psychology, 50(2), 229-238.

Gemar, M.C., Segal, Z.V., Sagrati, S., & Kennedy, S.J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. Journal of Abnormal Psychology, 110(2), 282-289.

Govan, C. L., & Williams, K. D. (2003). Reversing or eliminating IAT effects by changing the affective valence of the stimulus items. Journal of Experimental Social Psychology.

Greenwald, A. G. (2001). The Resting Parrot, the Dessert Stomach, and Other Perfectly Defensible Theories. Paper presented at the William J. McGuire Festschrift. April 21, 2001: Yale University.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4-27.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social Psychology, 79(6), 1022-1038.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464-1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. Zeitschrift für Experimentelle Psychologie, 48, 85-93.

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.

Kraut, R., Olson, J., Banaji, M.R., Bruckman, A., Cohen, J., & Couper, M., (2003). Psychological Research Online: Opportunities and Challenges. American Psychological Association.

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

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. Journal of Experimental Psychology: General.

Nosek, B. A. & Banaji, M. R. (2002). (At least) two factors moderate the relationship between implicit and explicit attitudes. In R.K. Ohme & M. Jarymowicz (Eds.), Natura Automatyzmow (pp. 49-56), Warszawa: WIP PAN & SWPS.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. Social Cognition, 19(6), 625-666.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). eResearch: Ethics, security, design, and control in psychological research on the Internet. Journal of Social Issues, 58(1), 161-176.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Harvesting intergroup implicit attitudes and beliefs from a demonstration Web site. Group Dynamics, 6(1), 101-115.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002c). Math = male, me = female, therefore math ? me. Journal of Personality and Social Psychology, 83(1), 44-59.

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

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

Wilson, T.D., Lindsey, S. & Schooler, T.Y. (2000). A model of dual attitudes. Psychological Review, 107(1), 101-126.

Table 1. Summary of data collection dates, sample size, and mean effects for the 11 Implicit Association Tests (IATs) comprising the data samples for the Studies 1-5.

	Task	Start date for sample	End date for sample	Total N	Excluded N	IAT D mean effect	IAT D SD	Effect size (d)
Yale IAT Web site	Bush-Gore attitude	11/15/01	10/29/02	7639	265	.03	.64	.05
	Black-White attitude	3/18/02	10/29/02	21925	540	.52	.51	1.02
	Gender-Science stereotype	4/30/02	10/29/02	11911	341	.42	.49	.86
	Old-Young attitude	3/18/02	10/29/02	12574	294	.62	.46	1.35
Tolerance.org Web site	Asian-White stereotype	11/15/01	6/17/02	4447	156	.37	.55	.66
	Black-White attitude	11/15/01	6/17/02	17050	309	.45	.54	.84
	Old-Young attitude	11/15/01	6/17/02	7186	182	.48	.45	1.07
	Dark-Light attitude	11/15/01	6/17/02	5254	124	.35	.48	.73
	Gender-Science stereotype	11/15/01	6/17/02	12269	377	.46	.49	.93
	Native-White stereotype	5/9/02	10/29/02	4495	149	.20	.60	.34
	Gay-Straight attitude	3/18/02	10/29/02	27220	960	.44	.57	.78

Table 2. Effects of manipulating the number of exemplars representing each IAT category (Study 2)

			Effect Magnitude				Internal Consistency		Extraneous Influences		
Task	No. of stimuli in target categories	No. of stimuli in attribute categories	IAT D mean effect	IAT D SD	Effect size (d)	Implicit-Explicit Corr	Practice-Test Corr	Attribute-Concept Corr	overall speed (r)	pairing order (r)	IAT experience (r)
Black-White attitude	6	8	.53	.53	1.01	.33	.55	.63	-.02	.07	-.10
	6	4	.54	.51	1.05	.35	.55	.63	-.02	.09	-.12
	6	2	.52	.53	.99	.34	.56	.63	-.02	.10	-.14
	6	1	.47	.46	1.01	.27	.43	.53	-.08	.04	-.07
Gender-Science stereotype	8	8	.47	.49	.96	.23	.53	.62	.04	.13	-.11
	4	4	.48	.49	.97	.24	.54	.62	.01	.21	-.09
	2	2	.43	.51	.85	.24	.53	.63	.05	.15	-.11
	1	1	.26	.43	.60	.17	.33	.44	.02	.03	-.09
Old-Young attitude	6	8	.64	.45	1.43	.15	.51	.59	.11	.07	-.22
	2	8	.60	.47	1.29	.11	.52	.60	.12	.19	-.19

Table 3. Effects of altering the exemplars representing a category for three tasks (Black/White, Gender-Science, Old/Young) (Study 3)

		Effect Magnitude				Internal Consistency		Extraneous Influences		
Task	Exemplar conditions	IAT D mean effect	IAT D SD	Effect size (d)	Implicit-Explicit Corr	Practice – Test Corr	Attribute-Concept Corr	overall speed (r)	pairing order (r)	IAT experience (r)
Black-White attitude	4a	.53	.51	1.04	.35	.54	.61	-.01	.10	-.10
	4b	.54	.51	1.06	.35	.57	.64	-.03	.15	-.08
Gender-Science stereotype	4a	.43	.48	.90	.24	.52	.63	-.01	.13	-.09
	4b	.53	.50	1.05	.24	.56	.61	.04	.17	-.10
Old-Young attitude	2a	.64	.47	1.36	.13	.54	.61	.14	.19	-.19
	2b	.60	.46	1.29	.07	.50	.57	.05	.22	-.19
	2c	.57	.47	1.21	.13	.52	.62	.15	.18	-.19
Native-American stereotype	Pictures	.18	.55	.32	.17	.64	.71	.04	.15	-.04
	Words	.23	.64	.36	.15	.68	.76	-.08	.28	.00
Gay-Straight attitude	Gay men	.52	.55	.95	.48	.64	.67	-.01	-.07	-.03
	Lesbians	.36	.58	.63	.51	.66	.71	.07	-.04	-.04

Note: Letter designations refer to multiple versions of the task using subsets of full stimulus set, but the same number of exemplars per set.

Table 4. Effects of varying the order of implicit and explicit measures (Study 4)

	Implicit effect magnitude			Explicit effect magnitude				Internal Consistency		Extraneous Influences		
Task	IAT D mean effect	IAT D SD	Implicit Effect size (d)	Explicit mean effect	Explicit SD	Explicit Effect size (d)	Implicit-Explicit Corr	Implicit practice - test	Explicit Likert-Therm.	overall speed (r)	pairing order (r)	IAT experience (r)
Black-White attitude												
Implicit First, Explicit Second	.51	.51	1.01	.40	1.27	.31	.32	.53	.71	-.02	.13	-.09
Explicit First, Implicit Second	.52	.52	1.01	.39	1.33	.29	.34	.54	.70	-.03	.11	-.08
Gender-Science stereotype												
Implicit First, Explicit Second	.39	.48	.80	.91	.84	1.05	.25	.50	.64	.06	.16	-.11
Explicit First, Implicit Second	.45	.49	.91	.80	.80	.98	.21	.52	.63	.08	.11	-.09
Old-Young attitude												
Implicit First, Explicit Second	.65	.46	1.41	.46	1.33	.35	.12	.53	.61	.12	.14	-.21
Explicit First, Implicit Second	.61	.46	1.32	.42	1.34	.32	.14	.50	.62	.11	.13	-.20
Average Effects												
Implicit First, Explicit Second	.52	.48	1.07	.59	1.15	.57	.23	.52	.65	.05	.14	-.14
Explicit First, Implicit Second	.53	.49	1.08	.54	1.16	.53	.23	.52	.65	.05	.12	-.12

Table 5. Effects of varying the order of compatible and incompatible blocks (Study 5)

	Effect Magnitude					Extraneous Influences	
Task	IAT D mean effect	IAT D SD	Implicit Effect size (d)	Implicit-Explicit Corr	Implicit internal consistency	overall speed (r)	IAT experience (r)
Black-White attitude							
'Compatible' Block First	.45	.54	.84	.33	.56	-.09	-.03
'Incompatible' Block First	.46	.53	.87	.40	.59	.04	-.17
Old-Young attitude							
'Compatible' Block First	.53	.46	1.16	.18	.47	.01	-.16
'Incompatible' Block First	.44	.44	.99	.21	.45	.06	-.15
Gender-Science stereotype							
'Compatible' Block First	.56	.47	1.19	.20	.51	-.01	-.05
'Incompatible' Block First	.35	.48	.72	.25	.53	.02	-.05
Asian-White stereotype							
'Compatible' Block First	.39	.56	.70	.20	.60	.07	-.06
'Incompatible' Block First	.33	.54	.61	.18	.59	.10	.07
Dark-Light attitude							
'Compatible' Block First	.38	.47	.81	.30	.50	.08	-.15
'Incompatible' Block First	.33	.50	.66	.24	.50	-.11	-.01
Average effects							
'Compatible' Block First	.46	.50	.94	.24	.53	.01	-.09
'Incompatible' Block First	.38	.50	.77	.26	.53	.02	-.06

Figure 1. Predicted and actual zero-order correlations among IAT and explicit attitudes or stereotypes calculated relatively or separately for individual target concepts. The top two panels illustrate the predicted of relationships if the IAT cannot be analytically decomposed (left), and if the IAT can be decomposed into independent association strengths (right). The bottom four panels present the observed effects for Bush-Gore, Black-White, Gender-Science, and Old-Young measures. The two independent explicit lines for the idealized non-decomposable hypothesis are not different; they are slightly separated to make both visible. (Study 1)

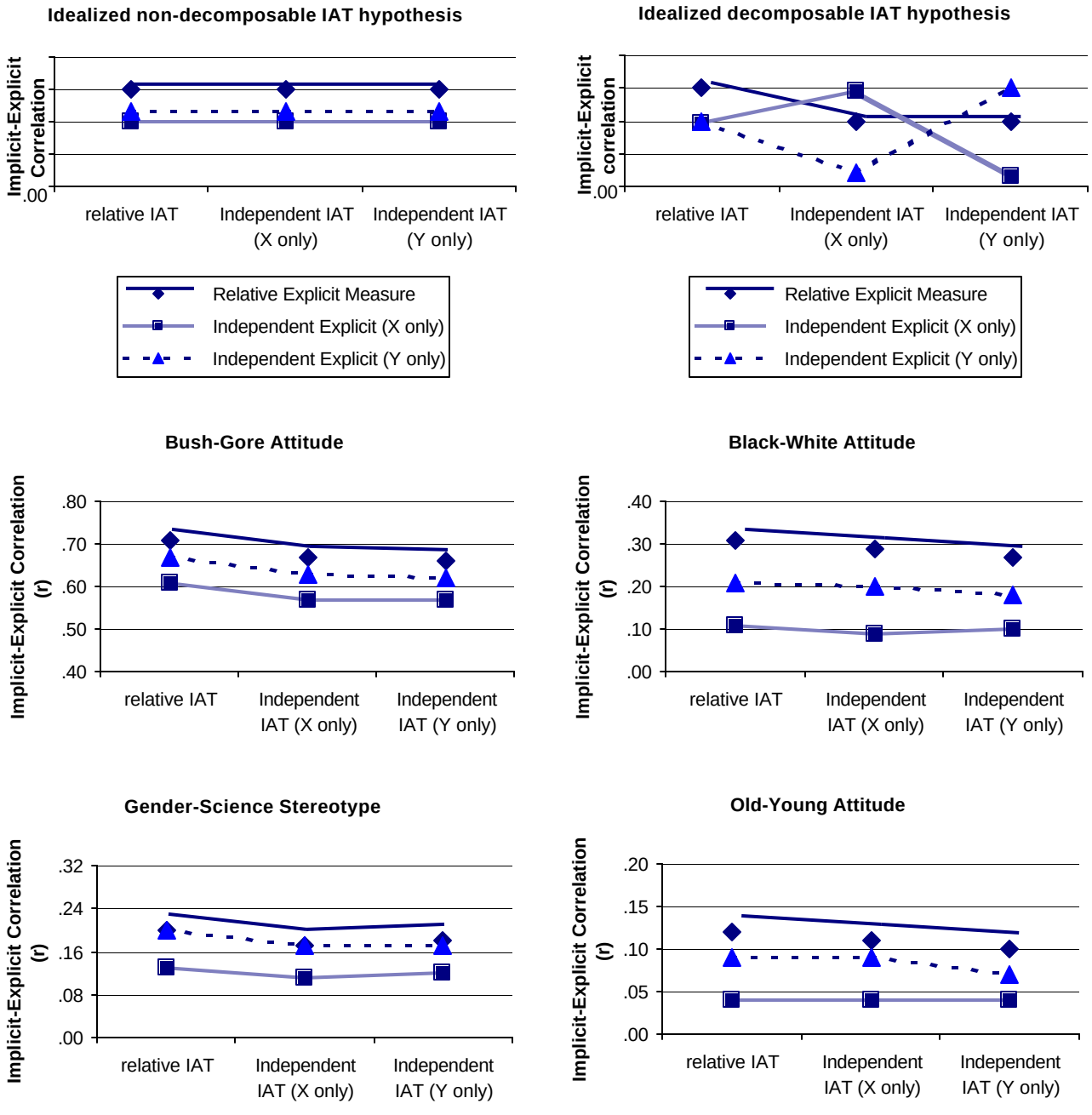
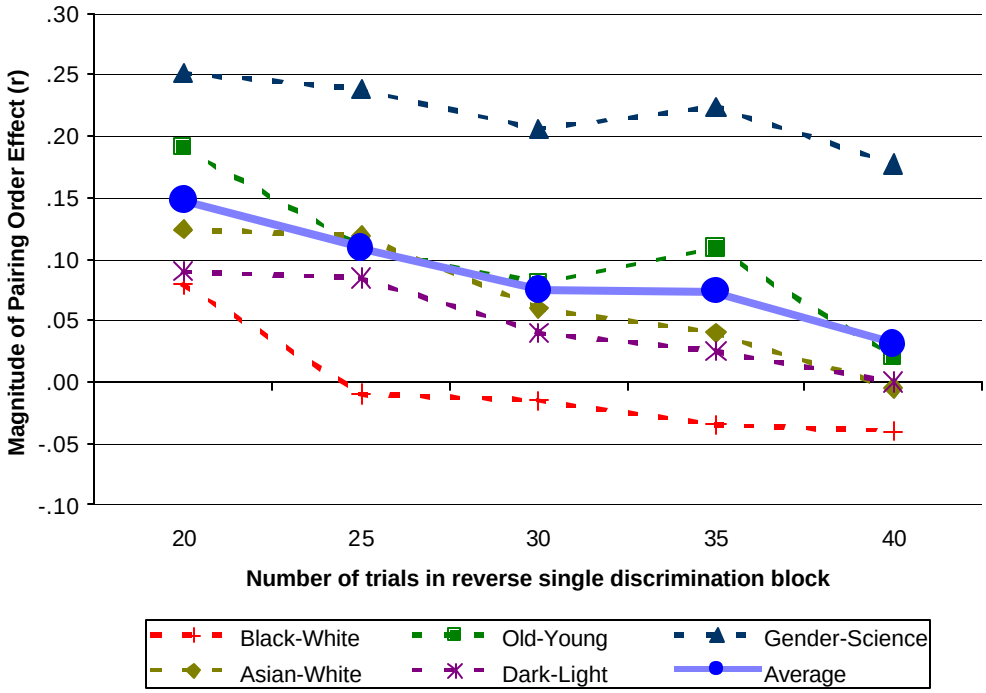


Figure 2. Magnitude of the pairing order effect by the number of response trials in the reverse single discrimination block for Black-White, Old-Young, Gender-Science, Asian-White, and Dark-Light IATs. A value of zero indicates the absence of a pairing order effect. (Study 5)



Footnotes

[1] The difference between Cohen's d and the present *D* measure is that the standard deviation in the denominator of d is a pooled within treatment standard deviation. The present *D* computes the standard deviation with the scores in both conditions, ignoring the condition membership of each score.

[2] The "Yale" Web site is now called Project Implicit, and resides at http://implicit.harvard.edu/.

[3] Similar analyses were conducted using the conventional scoring algorithm for the IAT described by Greenwald et al. (1998), and revealed effects consistent with all of those reported in this paper.

[4] The slightly higher implicit-explicit relationships for the standard 'relative' IAT calculation compared to the 'independent' calculations is attributable to the fact that the standard effect uses twice as many trials and is thus more reliable than the other two measures. When the difference in reliability is controlled, the three lines are horizontal for all four tasks.

[5] The single exemplar category label condition was introduced to the Black-White and Gender-Science tasks part way through the data collection after preliminary analysis of the effects suggested that it would be of interest as a comparison. The results reported in this paper include data from both before and after this extra condition was included. Results and interpretations do not vary if only the data collected after including this last condition are examined.

[6] Additional conditions described in the previous study with only two stimuli per category for the Black-White and Gender-Science tasks were also run. For simplicity, those data are not reported, but the effects observed with those versions are consistent with the effects reported here.

[7] Though it is not relevant for the current purposes, for the word task, American places were deliberately selected to be names of Native origins (e.g., Missouri). Also, for the picture task, the name of the landmark also appeared below the image in case respondents did not recognize the image on first viewing. So, the picture task is more appropriately termed a picture+word task. In any case, stimulus modality was very different than the word task.

[8] Various pretests in our laboratories have been performed to examine the boundary conditions of stimulus selection for IAT effects. For example, we have conducted successful tests of IATs in which all stimuli were eliminated in favor of temporarily representing one category (Math) with X's as exemplars and the other category (Arts) with O's as exemplars to show that stimuli can, in principle, exist simply as placeholders for the superordinate category. Because the representativeness of stimuli of their superordinate categories can enhance the quality of IAT effects, we recommend against such applications in IAT use except for parameter testing.