Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity

Brian A. Nosek
University of Virginia

Anthony G. Greenwald
University of Washington

Mahzarin R. Banaji
Harvard University

The Implicit Association Test (IAT) assesses relative strengths of four associations involving two pairs of contrasted concepts (e.g., male-female and family-career). In four studies, analyses of data from 11 Web IATs, averaging 12,000 respondents per data set, supported the following conclusions: (a) sorting IAT trials into subsets does not yield conceptually distinct measures; (b) valid IAT measures can be produced using as few as two items to represent each concept; (c) there are conditions for which the administration order of IAT and self-report measures does not alter psychometric properties of either measure; and (d) a known extraneous effect of IAT task block order was sharply reduced by using extra practice trials. Together, these analyses provide additional construct validation for the IAT and suggest practical guidelines to users of the IAT.

Keywords: *implicit social cognition; Implicit Association Test; attitudes; Internet; methodology*

The resurgence of interest in unconscious mental processes may be attributed to the availability of new measurement tools. Measures of implicit cognition differ from self-report in that they can reveal mental associations without requiring an act of introspection (Banaji, 2001; Bargh, 1997; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995; Wilson, Lindsey, & Schooler, 2000). In recent years, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) has been used to study implicit social cognition in part because of its ease of implementation, large effect sizes, and relatively good reliability (Greenwald & Nosek, 2001). Despite its popularity, there are many issues of design, analysis, and interpretation of IAT effects that are not yet understood. Numerous laboratories are tackling varied issues such as the processes underlying IAT effects (Mierke & Klauer, 2003; Rothermund & Wentura, 2004), the effects of procedural features such as stimulus exemplars (De Houwer, 2001; Steffens & Plewe, 2001), situational and temporal stability (Blair, 2002; Schmukle & Egloff, in press), the relationship between implicit and explicit measures (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2004; Nosek, 2004), and maximizing effectiveness of scoring procedures (Greenwald, Nosek, & Banaji, 2003). In addition to understanding the nature of implicit social cognition, these investigations offer pragmatic recommendations for maximizing the effectiveness of IAT design. The present article continues this method improvement effort by examining four questions concerning IAT procedures and application.

With more than 120 papers using the IAT in print in the 6 years since its original publication, there exists a healthy degree of variation in the procedures employed across studies and laboratories. Such procedural varia-

Authors' Note: This research was supported by the National Institute of Mental Health (MH-41328, MH-01533, MH-57672, and MH-68447) and the National Science Foundation (SBR-9422241 and SBR-9709924). The authors are grateful to Scott Akalis, Claudiu Dimofte, Jeff Ebert, Kristin Lane, N. Sriram, and Eric Uhlmann for their comments. Correspondence concerning this article should be addressed to Brian Nosek, Department of Psychology, University of Virginia, Box 400400, Charlottesville, VA 22911; e-mail: nosek@virginia.edu.

PSPB, Vol. 31 No. 2, February 2005 166-180
DOI: 10.1177/0146167204271418

tions introduce innovative ways of thinking about the IAT's task features and interpretation and have introduced many improvements to the original design. Random procedural variability also has the potential to introduce risks such as reducing the burden of providing a rationale for methodological variations, of allowing the effects of arbitrary procedural decisions on resulting data to go unnoticed, and of reducing the comparability of results across studies. In this article, some common design parameters are examined directly to provide an empirical basis for making methodological decisions about one's research objectives with the IAT.

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 (Greenwald et al., 1998). The IAT procedure obliges respondents to identify stimulus items and categorize them into one of four superordinate categories. Association strengths are measured by comparing the speed of categorizing members of the superordinate categories in two different sorting conditions. For example, because the concepts "Old" and "Bad" tend to be more strongly associated 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 (or blocks), 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 that 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: Concept-attribute pairing 1. 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 (often referred to as the "practice" block). After a brief pause, they repeat it for a second block of 40 trials (often referred to as the "critical" block).

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 an "a" key response.

Step 5: Concept-attribute pairing 2. 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 Young and Good items to be categorized with one key and Old and Bad items to be categorized with the other key, the opposite association from the earlier block. Respondents sort stimulus items with this response assignment for 20 trials 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. It 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 (called *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 four questions:

Question 1: Can analytic methods separate the IAT's measure of relative association strength into two separate measures of association strength? All attitude assessment occurs in a particular context that influences the resulting response (Schwarz, Groves, & Schuman, 1998). The IAT is conceptually similar to an explicit measure that asks a question about one group or idea in comparison to another. Although the IAT is structured as a relative measure, for many research questions it is desirable to seek separate assessments of single concepts. For example, distinguishing ingroup favoritism (e.g., Democrats liking Democrats) from outgroup derogation (e.g.,

TABLE 1: Summary of Data Collection Dates, Sample Size, and Mean Effects for the 11 Implicit Association Tests (IATs) Comprising the Data Samples for Studies 1-4

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	7,639	265	.03	.64	.05
Black-White attitude	3/18/02	10/29/02	21,925	540	.52	.51	1.02
Gender-Science stereotype	4/30/02	10/29/02	11,911	341	.42	.49	.86
Old-Young attitude	3/18/02	10/29/02	12,574	294	.62	.46	1.35
Tolerance.org Web site							
Asian-White stereotype	11/15/01	6/17/02	4,447	156	.37	.55	.66
Black-White attitude	11/15/01	6/17/02	17,050	309	.45	.54	.84
Old-Young attitude	11/15/01	6/17/02	7,186	182	.48	.45	1.07
Dark-Light attitude	11/15/01	6/17/02	5,254	124	.35	.48	.73
Gender-Science stereotype	11/15/01	6/17/02	12,269	377	.46	.49	.93
Native-White stereotype	5/9/02	10/29/02	4,495	149	.20	.60	.34
Gay-Straight attitude	3/18/02	10/29/02	27,220	960	.44	.57	.78

Democrats disliking Republicans) requires that evaluations of the ingroup and outgroup be assessed separately. Some researchers have analyzed subsets of IAT trials with the aim of obtaining separate evaluations of two target concepts in the IAT, but it is unclear if the structure of the IAT permits this aim. Study 1 tested whether this analytic strategy is empirically justified.

Question 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, being determined in part by the ease of generating suitable exemplars. Study 2 examined the impact of varying the number of items used to represent categories. This item has the pragmatic benefit of resolving whether IAT measures are appropriate even when categories have few words or exemplars that could sensibly represent it (e.g., gay-straight).

Question 3: Does the order of presenting IAT and self-report measures affect the outcome of either measure? In studies that use both IAT and self-report measures, decisions must be made about the order of presenting study materials. In a meta-analysis of IAT studies, Hofmann et al. (2004) observed that the correlation between the IAT and self-report was stronger when self-report came first. However, that observation was based on comparisons between experiments, not from a comparison of conditions established through random assignment. Study 3 experimentally investigated the effects of measurement order.

Question 4: Can the unwanted influence of order of IAT performance blocks be reduced? From the very first uses of the IAT, it has been known that the IAT effect is usually biased toward indicating greater strength of associative pairings involved in the first of the two combined tasks (Greenwald & Nosek, 2001). For example, in an age IAT, participants who first sort Old with Bad and Young with Good and then sort the reverse configuration show a stronger indication of implicit preference for Young over Old than participants who first sorted Old with Good and Young with Bad. Study 4 tested a procedural change designed to reduce this extraneous influence.

Data Source and Treatment

Two Web sites[2] served as the data source for the four studies reported in this article. Completing a test and a related questionnaire required about 10 min of participants' time. Afterward, respondents received a summary of their task performance along with a summary of previous respondents' results. In addition to typical debriefing materials, the Web site provided information about implicit social cognition more generally and provided answers to frequently asked questions. (For further discussion of the use of Internet-based data for such purposes, see Greenwald et al. [2003]; for a more general discussion about the Internet as the source of data collection, see Kraut et al. [2004] and Nosek, Banaji, and Greenwald [2002a].)

The large samples used for the present research (*N*s range from 4,447-27,220; see Table 1) provided sufficient power to guarantee that virtually all inferential tests would yield statistically significant results. In fact, all 11 samples in this article had 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 in the report is on effect size rather than significance level.

Three types of effect sizes are reported in this article. 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 was used. Cohen (1988) established conventions for interpreting effects as small, moderate, or large of .2, .5, and .8 for d and of .1, .3, and .5 for both r and q. As a guideline for this article, r and q effects less than .05 and d effects less than .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 others, links from search engines, or word of mouth. Data were taken from four tasks available at the Yale IAT Web site and seven tasks available at http://tolerance.org between November 15, 2001, and October 29, 2002 (see Table 1), representing a significant proportion of the data collected during this time frame. These tasks were selected to be representative of the variety of IAT applications for measuring implicit attitudes and stereotypes. Approximately 90% of respondents completed the demographic items. Of those responding (a) 66% were female and 34% were male; (b) 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; (c) 60% were less than 25 years of age, 35% were between 25 and 50, and 4% were older than 50; (d) 29% had a high school diploma or less, 57% had some college or a bachelor's degree, and 14% had an advanced degree; (e) 89% were from the United States; 6% were from Britain, Canada, or Australia; and 5% were from other countries; and (f) 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 comprised the data sets.

Materials and Apparatus

IATs were presented via the Internet using Java and CGI technology. 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 avoiding dependence on the speed of the Internet connection for accuracy of measurement. Accuracy of 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 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 across trials.

Implicit measures. The IATs measured implicit attitudes and stereotypes of a diverse range of social targets. Each of the tasks is described in brief, specifying the construct being measured along with the target and attribute concepts used in the measure: (a) *Ethnic-national stereotype 1 (Native-White)*: Native American/White American target concepts represented by faces and the attributes American/Foreign represented by words and images; (b) *Ethnic-national stereotype 2 (Asian-White)*: European/Asian concepts represented by faces and the attributes American/Foreign represented by images; (c) *Race attitude (Black-White)*: African Americans/European Americans target concepts represented by faces and the attributes Good/Bad represented by words; (d) *Skin color attitude (Dark-Light)*: Dark-Skinned/Light-Skinned people category labels and target concepts represented by faces and the attributes Good/Bad represented by words; (e) *Sexual orientation attitude (Gay-Straight)*: Gay People/Straight People represented by words and images and the attributes Good/Bad represented by words; (f) *Gender-Academic Domain stereotype (Male-Female)*: Male/Female categories represented by words and the attributes Science/Liberal Arts represented by words; (g) *Age attitude (Old-Young)*: Old/Young categories represented by faces and Good/Bad represented by words; and (h) *Political candidate attitude (Gore-Bush)*: Al Gore/George W. Bush represented by names and faces and Good/Bad represented by words.

Three of the IATs (Old-Young, Gender-Science, Black-White) appeared at two demonstration Web sites. Summary data for each task are listed in Table 1. Calculation of the IAT effect followed procedures described by Greenwald et al. (2003) and contained the following main features: (a) error trials were removed and replaced with the mean of that performance block plus a penalty of 600 ms and (b) individual trial response latencies less than 400 ms were removed before analysis. Also, respondents for whom > 10% of the trial responses were less than 300 ms were removed from the analysis (see the "Excluded *N*" column in Table 1).[3] The standard deviation of all response trials was used to calculate IAT effects, even for analyses using only subsets of the task trials. This introduced the possibility of extreme scores. Therefore, an additional exclusion criterion of dropping IAT scores at least 6 *SD*s away from the mean (an absolute value > 4.5) was added to eliminate very extreme outliers. This correction resulted in the removal of approximately 0.4% of the 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. Yet, the very large *N*, greater sample diversity than most types of data collections, variety of measures (8 unique, 11 total tasks), use of experimental manipulations, high power, and focus on methodological questions make this data set useful for the current purposes.

Multiple participations by individual respondents. Visitors to the demonstration Web sites were encouraged to try as many of the tasks as they wished. 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, previous investigations of data from these Web sites indicated that the inclusion or exclusion of multiple task performances from single respondents did not influence reported effects (Nosek, Banaji, & Greenwald, 2002b). Furthermore, a preliminary survey item asked respondents to report how many IATs they had completed previously. This measure was used to test the effect of prior experience on IAT performance, a potential extraneous influence.

STUDY 1: CAN ANALYTIC METHODS SEPARATE THE IAT'S MEASURE OF RELATIVE ASSOCIATION STRENGTH INTO TWO SEPARATE 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 compared to the strength of Flower + Bad/Insect + Good associations. 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 occasional publications, researchers have sought to interpret a portion of data collected within the IAT as meaningfully revealing such association measures for such subsets of blocks (e.g., Baccus, Baldwin, & Packer, 2004; de Jong, Pasman, Kindt, & van den Hout, 2001; Gemar, Segal, Sagrati, & Kennedy, 2001).

Gemar et al. (2001), for example, 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. They calculated average latencies for Me + Good responses when they required a single response and separately calculated an average latency for Not Me + Bad responses that shared the alternate response. Likewise, in the other condition, they calculated separate average latencies for Me + Bad responses and Not Me + Good responses. This way, the authors interpreted the comparison of Me + Good and Me + Bad latencies as a measure of self-evaluation and the comparison of Not Me + Good and Not Me + Bad as a measure of other-evaluation.

This analytic approach has attractive face validity in that it appears reasonable to analyze data for only one attitude object to get an assessment of evaluations toward that attitude object. Despite its appeal, the comparative response format that is part of the IAT design may actually result in every response capturing a component of the relative comparison between attitude objects. Using a multitrait/multimethod approach (Campbell & Fiske, 1959), we examined whether analyzing subsets of the IAT can yield meaningful indicators of separate attitude measures of the two target concepts.

Multitrait/multimethod comparison of IAT and self-report measures. The attitude construct has been hypothesized to have distinct implicit and explicit components, the former linked more closely to indirect measures such as the IAT and the latter to self-report (Greenwald & Banaji, 1995; Wilson et al., 2000). The evidence suggests that implicit and explicit attitudes are related but distinct constructs (Cunningham, Nezlek, & Banaji, 2004; Cunningham, Preacher, & Banaji, 2001; Greenwald & Farnham, 2000; Nosek & Smyth, 2004) and multiple moderators predict the strength of that relationship (Nosek, 2004). That is, implicit and explicit measures appear to have shared and unique attitude components.

The fact that implicit and explicit attitudes are related presents an opportunity to examine the construct validity of the analytic decomposition strategy of the IAT using a comparison to self-reported attitudes. Following the logic of the classic multitrait/multimethod matrix (Campbell & Fiske, 1959), implicit-explicit correlations should be maximized when measuring attitudes toward the same attitude object. That is, given that there is a relationship between implicit and explicit attitudes, if the IAT can be decomposed into single-category attitudes, then those individual attitudes should be more strongly related to same-trait explicit measures than cross-trait explicit measures. Implicit-explicit correlations need not be high but the comparative magnitude of correlations should conform to whether implicit and explicit measures are assessing the same or different attitude objects.[4]

Method

Materials. Data for four tasks at the Yale Web site were used to test whether the trial-subset analytic method can extract separate assessments of association strength with the IAT (Bush-Gore, Black-White, Gender-Science, and Old-Young). Additional analyses of the other seven data sets appearing in this article directly replicate the effects reported for these four data sets.

Calculating measures of relative versus separate association strengths with the IAT. Three IAT scores were calculated for each task. Using the Election 2000 task as an example, separate scores for Bush and Gore attitudes were calculated. Calculation of the relative Bush-Gore IAT score followed the standard format—all trials were retained.

Calculation of the Bush IAT score used only trial latencies involving categorization of items referring to George Bush and evaluative terms sharing a response key with Bush. Likewise, calculating the Gore IAT score used only trial latencies involving categorization of items referring to Al Gore and evaluative terms sharing a response key with Gore.

Analysis Strategy

For each task, separate 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 measure for Bush versus Gore. If the IAT can be decomposed, then single-category IATs should show the strongest correspondence with 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, implicit-explicit correlations should be strongest when the measures are assessing the same attitude object. In the present case, there are three attitude objects being assessed implicitly and explicitly—attitudes toward Bush, attitudes toward Gore, and attitudes toward Bush relative to Gore. Whether the correlation between implicit and explicit attitudes is large or small, comparatively, the 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 because the latter two are different attitude objects. Likewise, comparatively, the 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, comparatively, the Bush-Gore IAT should show the strongest relationship with explicit Bush versus Gore attitudes and somewhat weaker relationships with the separate 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 separate and relative IAT measures will all be measuring the same attitude construct—a relative preference of Bush versus Gore. In this case, all three IAT calculations (Bush IAT, Gore IAT, Bush-Gore IAT) would show similarly strong relations with the relative Bush versus Gore explicit attitude measure because it is the same attitude object (measured explicitly) and somewhat weaker relationships with explicit Bush attitudes and explicit Gore attitudes because they are different attitude objects. In other words, if the IAT is not decomposable, then the pattern of implicit-explicit correlations will be constant across trial subsets even if only trial responses from a single target category are examined. The pattern of correlations for the nondecomposable hypothesis is presented in the top left panel of Figure 1.

Results and Discussion

The four lower panels in Figure 1 present the patterns of correlations for the four tasks between the single-category and relative explicit attitude measures and the single-category and relative IATs for each task. If the IAT can be decomposed into separate attitude measures, then correspondence should be maximized when the IAT score and self-report are measuring the same attitude object (pattern in top right panel). If the IAT is not decomposable, then implicit-explicit correspondence should not vary as a function of the subset of trials comprising the IAT scores (parallel lines as shown in top left panel). The pattern across tasks clearly matches the idealized, nondecomposable pattern. The correspondence between the IAT and criterion variables does not vary as a function of the match between measurement types.[5] All three versions of the IAT appear to be measuring a single construct, not three different constructs. This indicates that the IAT cannot be analytically decomposed into separate assessments of association strengths. In other words, each response trial in the IAT appears indicative of the IAT effect as a whole, not an effect of the individual category or exemplar abstracted from the context of the task.[6]

Conclusion. Study 1 provides evidence that analytic decomposition strategies do not enable separate assessments 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. Measurement of single-category attitudes may require using a different measurement tool designed for such purposes, such as the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001) or the Extrinsic Affective Simon Task (EAST; De Houwer, 2003).

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. On one hand, increasing the number of exemplars may assist in providing a more accurate repre-

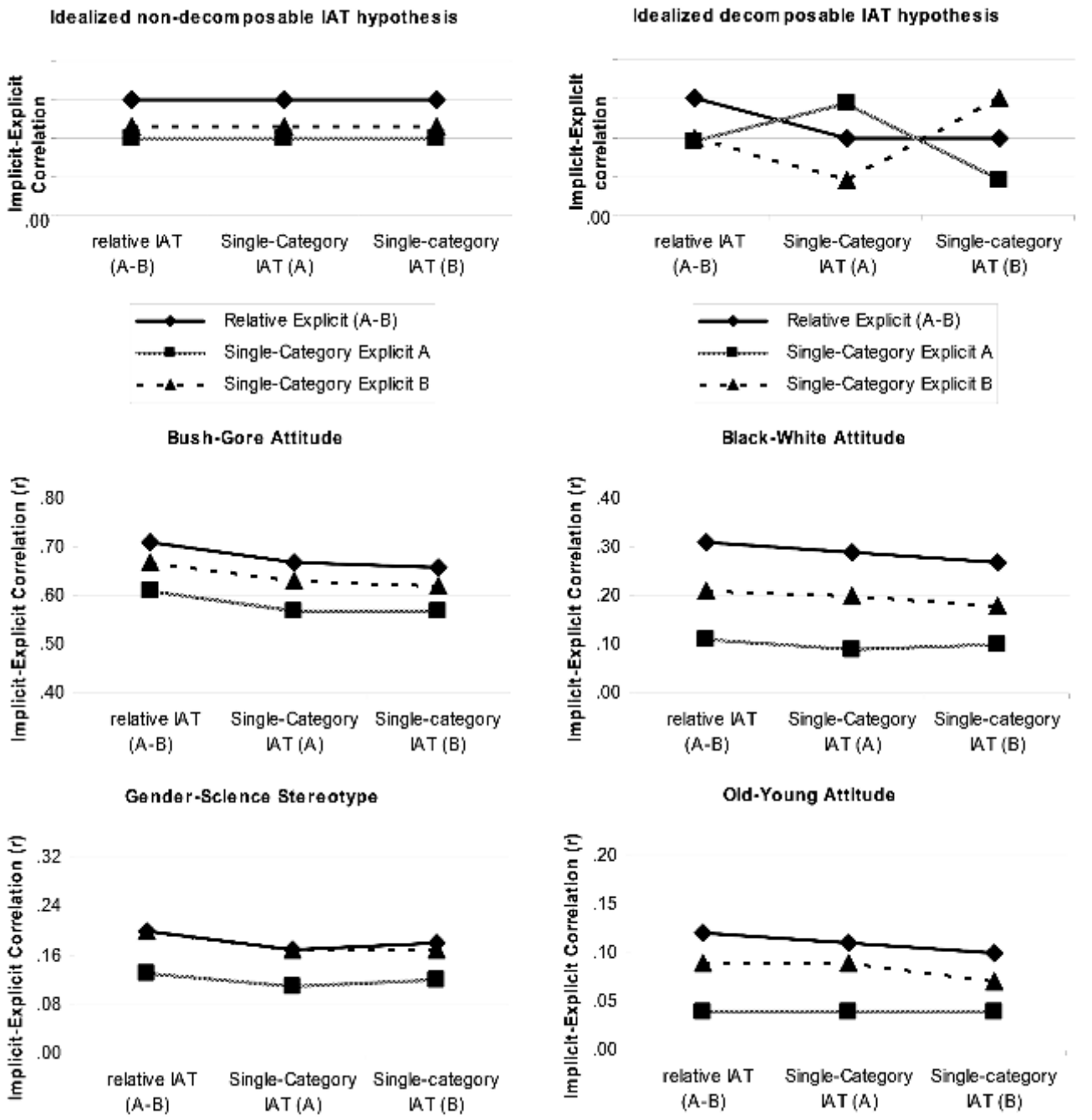

Figure 1 Predicted and actual zero-order correlations among IAT and explicit attitudes or stereotypes calculated relatively or separately for individual target concepts.

NOTE: The top two panels illustrate the predicted correlations of relationships if the Implicit Association Test (IAT) cannot be analytically decomposed (left) and if the IAT can be decomposed into separate association strengths (right). The bottom four panels present the observed effects for Bush-Gore, Black-White, Gender-Science, and Old-Young measures (Study 1).

sentation 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. Furthermore, if the IAT is psychometrically sound even in conditions with very few stimuli, then researchers will have more flexibility in IAT

design, especially for conditions in which very few sensible exemplars are available for the target concepts of interest.

In Study 2, we experimentally varied the number of exemplars 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. The experiment was conducted on three tasks providing replications across content domains (Black-White, Gender-Science, Old-Young from the Yale Web site). The Black-White task consisted of 6 exemplars (i.e., Black and White faces) for the Black and White categories and up to 8 exemplars (e.g., wonderful, terrible) for the Good and Bad attributes, for a total of 28 items. The Gender-Science task consisted of up to 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 up to 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 the task, 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 were held constant 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 (using the category labels "Good" and "Bad" as the stimulus items; 14 total).[7] For the Gender-Science task, the number of stimuli for both categories and attributes was manipulated across four conditions: 8 exemplars each (32 total stimuli), 4 exemplars each (16 total), 2 exemplars each (8 total), and 1 exemplar (using the category labels "Male," "Female," "Science," and "Liberal Arts" as the stimulus items; 4 total). For the Old-Young task, the 8 items representing the Good and Bad attribute categories were held constant 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), the exemplars were divided into subgroups such that, across subjects, all exemplars were used in equivalent amounts for each of the conditions.

Evaluation Criteria

Four criteria were used to evaluate the IAT for Studies 2 through 4: effect magnitudes, internal consistency, vulnerability to extraneous influences, and relation to self-report. These follow evaluation criteria that were established by Greenwald et al. (2003).

Effect magnitudes. Larger IAT *D* effect magnitudes for theoretically predicted effects were considered desirable.

Internal consistency. A measure of internal consistency was calculated by correlating IAT *D* effects for the first block of 20 trials of the two pairing conditions and the second block of 40 trials of the two pairing conditions. As an evaluation criterion, higher internal consistencies were considered desirable. In Study 4, internal consistency was calculated for self-reported attitudes and stereotypes by calculating the correlation between the comparative rating and warmth ratings described below.

Resistance to extraneous influences. Resistance to effects of three known extraneous influences on IAT effects—average response latency (McFarland & Crouch, 2002), task pairing order, and experience with the IAT—also served as evaluation criteria (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. Also, respondents were asked if they had previously performed an IAT. Following Greenwald et al.'s (2003) observation that experience effects were strongest between "0" and "1 or more," 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.

Self-report measures. Product-moment correlations between IAT effects and explicit measures served as a final criterion for evaluating IAT measures. As discussed previously, implicit and explicit attitudes appear to be distinct but related constructs (Cunningham, et al., 2001; Greenwald & Farnham, 2000; Nosek & Smyth, 2004), with implicit-explicit correlations ranging from near 0 to more than .7 (Nosek, 2004).

For any two conceptually related measures, reducing measurement error will necessarily increase the observed correlation between the measures to more closely approximate the "true" relationship between the variables. For example, height and weight are distinct

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*	*IAT Practice-Test Corr*	*IAT 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	8	8	.47	.49	.96	.23	.53	.62	.04	.13	–.11
stereotype	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

NOTE: IAT = Implicit Association Test.

but positively correlated constructs. Assessments of height and weight to the nearest millimeter and milligram will result in stronger height-weight correlations than would assessments to the nearest meter and kilogram. As long as the error between measures is uncorrelated, measurement error leads to underestimation of the true correlation between variables. Given the significant procedural differences between the IAT and self-report measures, it is unlikely that their correlation is due to method covariance (see also Nosek & Smyth, 2004). Therefore, the best IAT design will minimize measurement error and therefore maximize relations with known covariates, such as self-reported attitudes.

Associated with each IAT, respondents received a 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. Respondents rated their explicit preferences or stereotypes toward the target categories on a 5- or 9-point scale. For example, for the attitude measure for Old-Young preferences, the scale ranged from *I strongly prefer young people to old people* to *I strongly prefer old people to young people*. The two other explicit items were 11-point warmth rating thermometers in which the two target concepts were rated independently: "Please rate how warm or cold you feel toward the following groups (0 = *coldest feelings*, 5 = *neutral feelings*, 10 = *warmest feelings*)." The difference between the two warmth ratings served as the score. The two self-report measures were standardized and then averaged.

Results and Discussion

Table 2 presents effect magnitudes, relations to self-report, internal consistencies, and resistance to extraneous influence data for the three IATs separated by "number of exemplar" 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 magnitudes for the Black-White task showed only slight variation whether 8, 4, 2, or 1 exemplar(s) represented the categories "Good" and "Bad" (d range = .99 to 1.05). Also, the effect magnitude for the Old-Young task declined only slightly when the six exemplars representing each of the categories "Old" and "Young" were reduced to only two 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 one-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, 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 eight, four, or two 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 one-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 one-exemplar conditions, which were noticeably lower (average r for one-exemplar conditions = .43; average r for all other conditions = .58).

Varying the number of stimuli representing each category had little effect on three known extraneous influences of IAT effects. Although effects of pairing order varied somewhat depending on the number of stimuli,

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 decreased 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, although weaker, IAT effects emerged. IATs with only two items representing each attribute and category showed effects only slightly less robust than those using eight items per attribute and category.

Another interest regarding stimulus exemplars is the extent to which their semantic and evaluative properties influence IAT effects. Results from an additional study using these data sets suggested that using different exemplars from a homogeneous set of items does not influence IAT effects, but exemplars that influence the construal of the category labels do affect IAT scores.[8] This reinforces a similar point made elsewhere regarding the importance of selecting appropriate exemplars to represent the attributes and categories of direct interest (Govan & Williams, 2004; Mitchell, Nosek, & Banaji, 2003; Steffens & Plewe, 2001). Coupled with the present findings, this suggests that, all else being equal, selecting a small number of exemplars that are excellent representations of the target category will lead to better construct validity than selecting a large number of exemplars that are weak representations of the target category.

STUDY 3: DOES THE ORDER OF PRESENTING IAT AND SELF-REPORT MEASURES AFFECT THE OUTCOME OF EITHER MEASURE?

The preceding studies focused on features of the IAT procedure itself. However, IATs are frequently administered alongside other measures such as questionnaires. Does completing the IAT and perhaps gaining insight that one is showing a bias in a particular direction influence responding on a subsequent explicit measure of preference or stereotype? Alternatively, does increasing the accessibility of explicit attitudes or beliefs by administering explicit measures before implicit measures alter the effects observed on the IAT? In an interesting meta-analysis of the relationship between implicit and explicit attitudes, Hofmann et al. (2004) observed a small difference in implicit-explicit correlation when the self-reported measures preceded the IAT (K = 25, r = .24) than when the IAT preceded self-report (K = 40, r = .17). A limitation of this observation is that it was correlational, based on a between-studies comparison, rather than an experimental manipulation. Because researchers self-selected whether to use one order or the other, it is possible that choice of order was influenced by features of the attitude objects (e.g., perceived vulnerability to self-presentation). Furthermore, whereas measurement order may have influence under some conditions, that influence may not be consistent. In Study 3, we experimentally manipulated the order of IAT and self-report measurement to examine the influence of order on both measures.

Method

Measurement order was manipulated between subjects in three independent tasks (Black-White, Old-Young, and Gender-Science from the Yale Web site). After respondents selected a task, they were randomly assigned to receive the IAT or self-report measure first. Procedures, measures, and evaluation criteria mirrored those presented in Study 2.

Results and Discussion

Table 3 presents IAT and self-report effect magnitudes, relations between IAT and self-report, internal consistencies, and relations between IAT and extraneous factors for three tasks (Black-White, Gender-Science, Old-Young) separated by the order of presentation. The average effects across evaluation criteria reveal that the order of presentation of implicit and explicit materials had a minimal effect on both measures. In fact, average effects were near zero for all of the evaluation criteria, including IAT effect magnitudes (d = .02), explicit effect magnitudes (d = .06), implicit-explicit correlations (q = .00), implicit internal consistency (q = .00), explicit internal consistency (q = .00), and resistance to extraneous effects of overall speed (q = .00), task pairing order (q = .02), and experience with the IAT (q = .02). While showing little or no effects due to order of measures on average, small influences were observed on the magnitude of IAT (d = –.12) and self-report (d = .14) measures of Gender-Science stereotypes. The conclusion suggested by the current evidence is that the order of measurement does not have a strong influence on IAT or self-report for any of the evaluation criteria.

Despite the consistent lack of influence of order on the IAT and self-report, the specific nature of this measurement context may reduce the generality of this conclusion. 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 order would be observed if the target concepts were not known until

TABLE 3: Effects of Varying the Order of Implicit and Explicit Measures (Study 3)

	Implicit Effect Magnitude			*Explicit Effect Magnitude*				*Internal Consistency*		*Extraneous Influences*		
Task	*IAT D Mean Effect*	*IAT D* SD	*IAT Effect Size (d)*	*Explicit Mean Effect*	*Explicit* SD	*Explicit Effect Size (d)*	*Implicit-Explicit Corr*	*IAT Practice-Test*	*Explicit Reports*	*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

NOTE: IAT = Implicit Association Test.

measurement commenced. Reports from another large data set ($N > 11{,}000$), however, suggest that this is not the case. Nosek (2004) randomly assigned Web respondents to one of 57 different attitude object pairs and to measurement order. The order of IAT and self-report had no appreciable effect on the correspondence between implicit and explicit attitudes (Nosek, 2004). So, foreknowledge (or lack thereof) does not appear to introduce an effect of measurement order.

A second potentially important feature of the present study context is that the implicit and explicit measures were straightforward. Explicit measures consisted of 10 or fewer items about the attitude object and 8 or fewer items measuring demographic characteristics (e.g., age, sex, and 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 (Fazio, 1995) or by some other psychological mechanism. Even so, the study reported by Nosek (2004) had 25 to 29 self-report questions about the attitude object and still did not reveal an effect of measurement order.

There are, however, experimentally demonstrated effects of measurement order suggesting that, under some conditions, order will matter. For example, in one study, Bosson, Swann, and Pennebaker (2000) reported an effect of order between implicit and explicit measures. That study involved eight implicit and four explicit assessments in a single session. Also, demonstrations that implicit associations are influenced by previously performed tasks, contextual cues, and the immediate situation show the relevance of prior or current events on the IAT (Blair, 2002; Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Lowrey, Hardin, & Sinclair, 2001; Mitchell et al., 2003). IAT scores may be influenced by preceding self-report when the report changes the context of measurement or alters the underlying associations.

Finally, although speculative, 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 as relevant information for generating a self-reported preference when implicit measures are first.

Conclusion. The present study suggests that the effects of measurement order may not be a consistent or reliable effect as suggested by the Hofmann et al. (2004) meta-analysis. Minimal effects of task order were observed in measurement contexts whether the target attitude objects were known in advance and when the self-report measure consisted of almost 30 attitude-relevant questions. Also, performing the IAT before self-report does not appear to induce reactance or assimilation tendencies in subsequent self-report. Even so, because some experimentally demonstrated order effects do exist, there must be moderators that will predict the influence of measurement order. Until those moderators are identified, researchers should carefully consider measurement order, especially for situations in which self-report or IAT measures are extensive or when the attitude objects are relatively novel, unstable, or likely to elicit ambivalent responses.

STUDY 4: CAN THE UNWANTED INFLUENCE OF ORDER OF IAT PERFORMANCE BLOCKS BE AVOIDED?

A persistent methodological artifact for the IAT is the order of task performance blocks (Greenwald & Nosek,

2001). For example, respondents who first complete the condition where Insects are paired with Bad and Flowers are paired with Good 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. Klauer and Mierke (2004) suggest that this effect may be due to after-effects of task-switching that are present in conditions that require the same response for nonassociated categories (incompatible) to a greater extent than conditions that require the same response for associated categories (compatible). Whatever its cause, this measurement artifact produces undesirable error in IAT measurement. In this section, we examine the effects of pairing order and test a methodological innovation to remove its influence 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 subsequently needing to replace it with a new response set, then extra practice with the new response set may reduce this effect. Our strategy to eliminate the order effect was to increase the number of trials in the fourth block in which the sorting instructions for target concepts has been reversed. In Study 4, we varied the number of trials in the fourth block by increments of 5 trials between 20 and 40 trials for five different tasks. We tested whether adding trials to this block was sufficient to make the magnitude of the IAT effect equivalent between pairing order conditions.

Method

Materials. The manipulation of practice trials was conducted on five different tasks (Old-Young, Asian-White, Gender-Science, Black-White, and Dark Skin-Light Skin) from those available at http://tolerance.org, providing multiple replications.

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

Results and Discussion

The impact of pairing order on effect magnitude, reliability, relationship with explicit attitudes and stereotypes, and resistance to extraneous influences. Initial tests showed that the experimental manipulation of practice trials only influenced the magnitude of IAT effects and not the other evaluation criteria: internal consistency, resistance to extraneous influence, and implicit-explicit correspondence (all ds and qs < .05, most near 0). Therefore, Table 4 summarizes the evaluation criterion by the IAT pairing order after collapsing across the between-subjects manipulation. The effect of the manipulation on effect magnitudes appears in Figure 2 and is discussed in the next section.

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 underrepresents its "typical" magnitude because the reported means collapse across the conditions designed to reduce its impact (see the 20 trials condition in Figure 2 for the "typical" effect of task order). Even so, larger effects tended to be observed when the compatible block of the task was performed first compared to second (average d = .18). However, the influence of pairing order on other evaluation criteria was weak. Performing the compatible task first was associated with nondifferentiated relations between implicit and explicit preferences (average q = .02), internal consistency (average q = .00), vulnerability to extraneous influences of overall speed (average q = –.01) and experience with the IAT (average q = –.03). This suggests that, across five independent content domains, one pairing order does not provide a better estimate of the underlying construct than the other. The impact of pairing order appears to have exclusive influence on the magnitude of IAT effects and not on its reliability, relations with self-report, 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

TABLE 4: Effects of Varying the Order of Compatible and Incompatible Blocks (Study 4)

	Effect Magnitude					*Extraneous Influences*	
Task	*IAT D Mean Effect*	*IAT D* SD	*IAT Effect Size (d)*	*Implicit-Explicit Corr*	*IAT 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

NOTE: IAT = Implicit Association Test.

number of trials in the reverse single discrimination practice block for five tasks and the unweighted average of the effects for 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.

Of interest, although the order effect was certainly reduced for the Gender-Science task (r = .25 down to r = .17, a 54% reduction in shared variance), it was not 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 whereas 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. At present, this hypothesis is speculative.

Conclusion. Results from Study 4 suggest that adding additional practice to the reversed single discrimination practice block will reduce pairing order effects, and even eliminate them. This constitutes an important improvement to construct validity because it reduces or removes one of the most persistent extraneous influences on IAT effects. Even with the added practice, the effect was not eliminated completely for the Gender-Science task. Until the moderating factor of this effect is identified, added practice trials and continued use of pairing order counterbalancing will ensure minimal influence of this irrelevant factor.

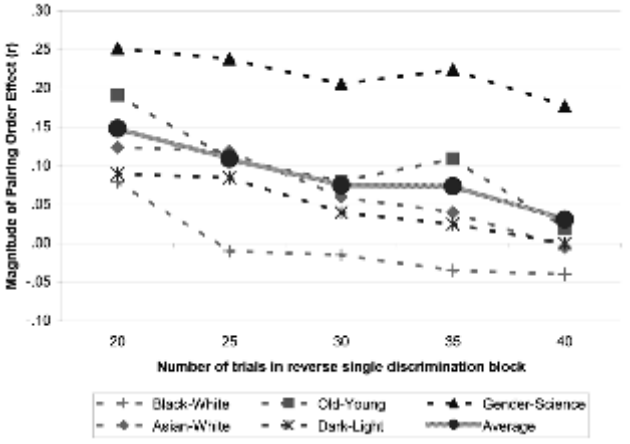

Figure 2 Magnitude of the pairing order effect by the number of response trials in the reverse single discrimination block for Black-White, Old-Young, Gender-Science, Asian-White, and Dark Skin-Light Skin Implicit Association Tests (IATs). A value of zero indicates the absence of a pairing order effect (Study 4).

GENERAL DISCUSSION

In four studies, we investigated methodological issues relevant to the design, analysis, and interpretation of the IAT. The results provide an empirical basis for informing decisions about procedural design in studies that use the task. The findings can be summarized as follows:

Question 1: Can analytic methods separate the IAT's measure of relative association strength into two separate measures of association strength? 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. Researchers interested in assessing associations with a single target concept should use a method designed for that purpose (e.g., De Houwer, 2003; Nosek & Banaji, 2001).

Question 2: Is there an optimal number of stimulus items per category in the IAT? 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 two stimuli per category. Decisions about the number of stimuli to use for an IAT can be based on pragmatic concerns, with at least four stimulus items per category appearing to be ideal but two 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.

Question 3: Does the order of IAT and self-report measures affect the outcome of either measure? In Study 3, little to no effect on magnitude of implicit and explicit measure means was observed as a function of order in which the implicit and explicit measures were presented. This contrasts with a recent meta-analysis (Hofmann et al., 2004) that was limited to between-study comparisons of task order. The results of the present study suggest that performing the IAT before self-report does not induce reactance or assimilation effects in subsequent self-report. And, coupled with supplementary analyses of another large data set (Nosek, 2004), the lack of measurement order effects cannot be attributed to self-selection of tasks or foreknowledge of content domain of the study. However, the generality of these observations may be constrained by evidence for measurement order effects when situational or contextual factors are altered (Blair, 2002; Bosson et al., 2000). Practical concern about the presentation order of implicit and explicit measures may be unnecessary when the measures are relatively short and simple and where responses to the target concepts are likely to be stable and unambivalent. Nevertheless, the cautious strategy of counterbalancing order of administration of measures may be soundest when there is no compelling reason to favor one order.

Question 4: Can the unwanted influence of order of IAT performance blocks be reduced? One of the most robust and well-documented extraneous influences on the IAT is the order of task performance blocks. In Study 4, we reproduced this widely observed effect of pairing order and provided evidence that a simple procedural change can dramatically reduce its influence. Doubling the number of trials in the reverse single-discrimination block of trials from 20 to 40 (in Step 4 of the five-step IAT procedure) reduced the overall impact of task order to $r = .03$. This procedural change has the desirable consequence of minimizing an often-significant extraneous influence on IAT effects. Of importance, for one task, Gender-Science stereotype, the order effect did decline with this procedural adjustment, but it did not disappear. We speculated that pairing order effects are more robust with IATs using lexical stimuli exclusively.

Conclusion

The necessary link between theory and method in science makes the rigorous examination of method of critical importance for the advancement of theory. Pragmatically, attention to methodological questions can increase efficiency with which the collective research enterprise can focus on theoretical questions. In this article, four 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 these results will accelerate theoretical exploration of implicit social cognition.

NOTES

1. The difference between Cohen's *d* and the Implicit Association Test (IAT) *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. These sites were previously located at http://www.yale.edu/implicit/ and http://tolerance.org/. At the time of writing this article, those two sites have been replaced by http://implicit.harvard.edu/.

3. A bug in recording some of the latencies in error trial responses required the error replacement strategy discussed by Greenwald, Nosek, and Banaji (2003) rather than retaining the error trial latencies as is.

4. In fact, justification of a conceptual distinction between implicit and explicit attitudes requires that implicit and explicit measures each capture distinct, attitude-relevant variation (Banaji, 2001). Such a conceptual distinction does not, however, require that implicit and explicit attitudes be completely unrelated. The fact that implicit and explicit attitudes are related merely eliminates the most extreme form of dissociation—that they are exclusive constructs.

5. The slightly higher implicit-explicit relationships for the standard relative IAT calculation compared to the single-category IAT calculations is attributable to the fact that the relative IAT score 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.

6. A reviewer suggested that the decomposition strategy may work for self-esteem measures (similar to those used by researchers previously; Gemar, Segal, Sagrati, & Kennedy, 2001) even though it did not

for various attitude object pairs that we tested. We tested this possibility with a large sample of self-esteem IAT data reported by Banaji and Nosek (2004; $N = 6{,}229$). That analysis replicated effects reported here.

7. The single exemplar category label condition was introduced to the Black-White and Gender-Science tasks partway 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 article include data from both before and after this extra condition was included. Results and interpretations were the same using only the data collected after including this last condition.

8. A report of this follow-up study is available at http://briannosek.com.

REFERENCES

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. *Psychological Science, 15*(7), 498-502.

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.

Banaji, M. R., & Nosek, B. A. (2004). *Implicit racial identity, attitude, and self-esteem.* Unpublished manuscript, Harvard University.

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: Lawrence Erlbaum.

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 implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631-643.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin, 30*, 1332-1346.

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 Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50*(2), 77-85.

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. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength* (pp. 247-282). Mahwah, NJ: Lawrence Erlbaum.

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. (2004). Reversing or eliminating IAT effects by changing the affective valence of the stimulus items. *Journal of Experimental Social Psychology, 40*(3), 357-365.

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, 85*(2), 197-216.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2004). *A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures.* Unpublished manuscript, University of Trier, Germany.

Klauer, K. C., & Mierke, J. (2004). *Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the IAT effect.* Unpublished manuscript, Rheinische Friedrich-Wilhelms-Universitat, Bonn.

Kraut, R., Olson, J., Banaji, M. R., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Opportunities and challenges. *American Psychologist, 59*(2), 105-117.

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

McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition, 20*(6), 483-510.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*(6), 1180-1192.

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*(3), 455-469.

Nosek, B. A. (2004). *Moderators of the relationship between implicit and explicit attitudes.* Unpublished manuscript, University of Virginia, Charlottesville.

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., & Smyth, F. (2004). *Implicit and explicit attitudes are related but distinct constructs.* Unpublished manuscript, University of Virginia, Charlottesville.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*(2), 139-165.

Schmukle, S. C., & Egloff, B. (in press). Does the Implicit Association Test for assessing anxiety measure trait and state variance? *European Journal of Personality.*

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vol. 1* (4th ed., pp. 143-179). Boston: McGraw-Hill.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift Fuer Experimentelle Psychologie, 48*(2), 123-134.

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

Received February 19, 2004
Revision accepted May 17, 2004