Journal of Personality and Social Psychology
2008, Vol. 94, No. 1, 16–31

0022-3514/08/$12.00 DOI: 10.1037/0022-3514.94.1.16

Why Do Implicit and Explicit Attitude Tests Diverge? The Role of Structural Fit

B. Keith Payne
University of North Carolina at Chapel Hill

Melissa A. Burkley
Oklahoma State University

Mark B. Stokes
University of North Carolina at Chapel Hill

Implicit and explicit attitude tests are often weakly correlated, leading some theorists to conclude that implicit and explicit cognition are independent. Popular implicit and explicit tests, however, differ in many ways beyond implicit and explicit cognition. The authors examined in 4 studies whether correlations between implicit and explicit tests were influenced by the similarity in task demands (i.e., structural fit) and, hence, the processes engaged by each test. Using an affect misattribution procedure, they systematically varied the structural fit of implicit and explicit tests of racial attitudes. As test formats became more similar, the implicit–explicit correlation increased until it became higher than in most previous research. When tests differ in structure, they may underestimate the relationship between implicit and explicit cognition. The authors propose a solution that uses procedures to maximize structural fit.

Keywords: implicit, attitude, automatic, measurement, prejudice

Implicit tests have been compared with such revolutionary inventions as the telescope and the microscope. The hope is that implicit tests, too, can make clear what is invisible to the naked eye. In many studies, implicit tests have created images of attitudes and beliefs that look very different from those reported on questionnaires. This kind of divergence is especially common for tests of racial attitudes and stereotypes. But what exactly does it mean when a person reports one attitude yet scores differently on an implicit test of race bias? The answer to that question is controversial, but it is important. It will shape not only theories of attitudes and stereotypes, but also the way that men and women taking the tests understand their own minds (see Arkes & Tetlock, 2004, and commentaries; Blanton & Jaccard, 2006, and commentaries).

In a typical study of this sort, a sample of research volunteers is compared on two tests of racial attitudes. One test is explicit, asking them to report their attitudes on a questionnaire. The other test is implicit. Rather than asking for self-report, it uses performance on another task to reveal attitudes. Readers familiar with implicit social-cognition research over the past decade will have no trouble predicting that in this kind of study, the two measures will likely diverge, capturing two very different snapshots of racial attitudes (Fazio & Olson, 2003; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005).

But why do implicit and explicit measures diverge? One view is that the two kinds of measures reflect separate attitude representations (Devine, 1989; Wilson, Lindsey, & Schooler, 2000). By this account, people hold multiple attitudes toward a topic at the same time. When attitudes change, a new attitude is layered on top of older attitudes. When people introspect they report the most contemporary attitudes, but the ruins of older layers can be unearthed by probing deeper, using implicit tests.

A different view is that a lack of correlation between measures does not turn up separate attitudes at all. Instead, the two kinds of measures allow people to edit their responses to different degrees. From this point of view, measuring implicit responses is less like an archeological dig and more like fishing in a river. Implicit tests tap attitudes upstream, but explicit tests catch what flows downstream, muddied in the editing for public report (Fazio, Jackson, Dunton, & Williams, 1995; Nier, 2005).

Both perspectives assume that the chief reason for the implicit–explicit divide can be found in the distinction between "implicitness" and "explicitness." Either implicit measures tap something unconscious and explicit measures tap something conscious, or implicit measures tap automatic responses and explicit measures tap intentionally edited responses. It may seem obvious that the principal difference between implicit and explicit tests is that one is implicit and the other is explicit. To see what other possibilities exist, it helps to shift perspectives and ask how implicit and explicit tests differ beyond implicit and explicit cognition. We propose that, independent of differences in underlying cognitive processes, when implicit and explicit tests have radically different structures, they will correlate with each other only weakly. But

B. Keith Payne and Mark B. Stokes, Department of Psychology, University of North Carolina at Chapel Hill; Melissa A. Burkley, Department of Psychology, Oklahoma State University.

This research was supported by the National Science Foundation Grant 0615478 to B. Keith Payne.

Correspondence concerning this article should be addressed to B. Keith Payne, Department of Psychology, University of North Carolina at Chapel Hill, Campus Box 3270, Chapel Hill, NC 27599. E-mail: Payne@unc.edu

when they have similar structures, they will show much greater agreement.

If this analysis is correct, then it has two important implications for the study of implicit social cognition. First, it would call into question whether the many null or weak implicit–explicit correlations that have been reported should be interpreted as evidence that underlying implicit and explicit cognitions are independent. Second, to draw conclusions about implicit and explicit thought processes from divergence between implicit and explicit tests, one must first equate the tests on extraneous differences while systematically varying the differences of interest. Our aim is not, therefore, to test whether the archeological metaphor or the river metaphor is correct. Our aim is, instead, to show that poor structural fit creates a stumbling block for investigating such theories with implicit tests and to propose a way around this obstacle.

Structural Fit Between Implicit and Explicit Tests

By the *structure* of a test, we mean the parts that make it up and how they work together to measure attitudes. Most explicit attitude tests share several structural elements. The items are usually verbal statements. For example, an item on the Likert-style Attitudes Toward Blacks Scale (ATB) reads, "Racial integration (of schools, businesses, residences, etc.) has benefited both Blacks and Whites" (Brigham, 1993, p. 1942). In a semantic differential, participants might be asked to rate racial groups on traits such as "pleasant," "aggressive," and "friendly." And on a feeling thermometer, participants might be presented with several racial groups and asked to rate their feelings toward each group from *very cold and unfavorable* to *very warm and favorable*. In each of these cases, participants read a verbal phrase or sentence. They must either retrieve a previously stored attitude from memory or construct a new evaluation in the moment. Finally, they must decide how to best express their response to each statement on a numerical scale.

If explicit attitude measures involve considering statements, evaluating one's response, and formulating it on a scale, implicit measures avoid all of these. Although the structures of implicit tests differ from one procedure to another, certain commonalities are clear. Complex propositions are replaced with simple words or pictures. In the implicit association test, for example, words or pictures denote the target items to be evaluated (Greenwald, McGhee, & Schwartz, 1998). Participants are asked to classify that item using four categories mapped onto only two overlapping response keys (e.g., White or good, Black or bad, White or bad, Black or good). In *evaluative priming* (Fazio et al., 1995), a prime word or picture is flashed briefly before a target word or picture. The target item is then evaluated as "good" or "bad." Similar items are presented for many other kinds of tasks (e.g., De Houwer, 2003; Wittenbrink, Judd, & Park, 1997).

When presented with a word or a picture, implicit test takers are asked to simply evaluate it, categorize it, or decide whether it is a word. This does not typically require the formulation of any opinion, as there is usually a correct answer (e.g., *death* is bad). In response-latency tasks, which make up the bulk of implicit measurement, the content of the response is irrelevant (incorrect answers are typically excluded). The measure of interest is the time it takes to register a response.

The list of differences described here between implicit and explicit tests is not intended to be exhaustive; these examples merely highlight some key ways that implicit and explicit tasks differ. They include the stimuli presented (e.g., propositions vs. simple words or pictures), the level of abstractness of the judgments to be made (e.g., broad social opinions vs. concrete classifications), and the metric in which responses are measured (e.g., numerical scales vs. response latencies). It is an important fact that none of these differences is inherently related to consciousness or unconsciousness, automaticity or voluntary control. Instead, they are incidental properties that are confounded with the implicit–explicit distinction as it has been instantiated in popular methods.

Are these methodological differences important for understanding implicit–explicit correlations? The issue calls to mind earlier debates about the predictive value of attitudes in general. Faced with many failures to detect relationships between attitudes and behaviors (Wicker, 1969), attitude theorists uncovered a number of moderating factors that determined when attitudes and behaviors are likely to be related and when they are not. One key factor was *conceptual correspondence* (Ajzen & Fishbein, 1977).

When attitudes and behaviors are measured at the same level of abstractness and with the same degree of specificity, they are said to be conceptually correspondent. Under these conditions, attitudes and behavior tend to be related. Attitudes toward good health, for example, are not strongly related to how often a person jogs, but attitudes toward jogging are more likely to be related. A recent review supports the notion that implicit and explicit attitude measures are more likely to be related when they are conceptually correspondent (Hofmann et al., 2005). Of course, conceptual correspondence is only one aspect of the structural differences we describe between implicit and explicit tests. Differences such as reaction times versus Likert scales are important parts of a test's structure but are unrelated to conceptual correspondence. For a more inclusive and accurate description of our purposes, we refer to the degree of methodological similarity between different tests as *structural fit*.

The attitude–behavior relationship is not the only field of study in which issues of test structure have proved important. Early studies of implicit memory compared implicit and explicit memory tasks that differed in many ways. For example, recall and recognition tasks measured explicit memory. In contrast, implicit memory was measured with a range of tasks, including word-fragment completion, word identification, and lexical decision (Jacoby & Dallas, 1981; Tulving, Schacter, & Stark, 1982; Warrington & Weiskrantz, 1974). Researchers using this approach found many variables that selectively influenced one kind of test but not the other. These dissociations, however, only begged more questions about their underlying reasons. Given all the structural differences between, for example, a recall test and a lexical-decision test, it was not clear whether a dissociation reflected implicit versus explicit forms of memory or other differences in the operations that each task requires (Roediger, 1990). Ambiguities in how to interpret implicit tests led Schacter, Bowers, and Booker (1989) to propose a principle they called the *retrieval intentionality criterion*.

The retrieval intentionality criterion says that to isolate implicit and explicit forms of memory in a way that is empirically verifiable, implicit and explicit memory should be measured in a way that holds everything about the memory tests constant except the intention to remember. The intention to remember is then manipulated. For example, rather than comparing cued recall to lexical

decision, a study should present the same word-fragment cues for both implicit and explicit tests. In the implicit test, instructions should require participants to complete them in a way that does not refer back to a studied event (e.g., "complete the fragment with the first word completion that comes to mind"). In the explicit test, participants should complete the same items under instructions to remember the previous event (e.g., "complete the fragment with a word that you studied").

When some variable affects one kind of test but not the other, the dissociation provides evidence for a selective effect on intentional versus unintentional uses of memory. This approach links the operational definition of implicit memory to its conceptual definition, because implicit memory is defined as an effect of past experience that does not require the intent to remember. Jacoby (1991) later provided a more conservative definition in which participants in the implicit test condition are told to complete the fragment with a word that they did *not* study. In this case, participants would produce a studied word only if it came to mind but was *not* consciously remembered. This exclusion instruction defines implicit memory as an effect of past experience that influences performance despite a conscious intention to the contrary.

The retrieval intentionality criterion is based on a fundamental principle of experimental design: Isolating a particular variable requires that all other variables be held constant. Doing otherwise allows a confound in the design. Although the retrieval intentionality criterion soon became a gold standard for implicit memory research, studies of implicit attitudes have not followed the same route. Research on the attitude–behavior link and on implicit memory have both shown how important it can be to hold extraneous factors constant. The studies reported here explore how the relationships between implicit and explicit attitude tests change when the test structures are matched.

Overview of the Present Research

To test whether structural fit influenced implicit–explicit correlations, we manipulated how well these features were equated. We reasoned that if extraneous structural differences led to underestimated correlations, then the more closely tests were equated on these features, the stronger the correlations would be. It is difficult to use response-latency methods with this approach, because the structure of the test is essential to the function of the test. It would be impossible to overcome problems such as comparing response latencies to Likert scales without changing the nature of the tasks. Several authors have observed, in various ways, that methodological differences might reduce the relationships between reaction-time tests and self-report tests (e.g., Hofmann et al., 2005; Kawakami & Dovidio, 2001; Wittenbrink et al., 1997). But the question is far from settled, because there has generally been no alternative available. As a result, there has been no way to gauge the effect of these differences and no way to know what the relationships would be in their absence. In this article, we gauge how much of an effect structural fit may have, and we propose an approach for greatly reducing extraneous differences.

An important step in understanding implicit–explicit correlations is the finding that a latent-variable approach can greatly increase the implicit–explicit correlation by removing measurement error (Cunningham, Preacher, & Banaji, 2001). The question addressed in the present research, however, is different. Latent-variable analyses can estimate what relationships would be in the absence of random error, but the method differences on which we focus represent systematic method differences that cannot be removed with a latent-variable approach. Instead, multiple measures are needed to compare the systematic influence of different methods (Bagozzi & Yi, 1991; Campbell & Fiske, 1959; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

To solve this problem, we took advantage of the recently developed *affect misattribution procedure* (AMP; Payne, Cheng, Govorun, & Stewart, 2005), because it does not rely on response latencies. Instead, it produces an implicit measure of attitudes in which the response metric is an evaluation. The AMP is an approach to implicit measurement that depends on evaluation of ambiguous items. When an ambiguous object, such as a Chinese pictograph, is preceded with a pleasant or unpleasant picture, the picture alters impressions of the pictograph (Murphy & Zajonc, 1993). People tend to misattribute their affective reaction from the prime picture to the target pictograph. As a result, participants asked to rate the pleasantness of the pictograph tend to rate it as more pleasant following a smiling face as compared with a frowning face. The measure of interest is not reaction time but the pictograph's rated pleasantness.

Participants showed strong misattribution effects even when directly warned to avoid any influence from the prime photos (Payne, Cheng, et al., 2005). A key aspect of the AMP is that participants are warned specifically that the prime photos may bias their evaluations of pictographs, and they are instructed that their task is to avoid any influence from the photos. Providing such a warning sets intentional response strategies in opposition to the automatic influence of the primes (an exclusion instruction; Jacoby, 1991). If participants respond as intended, they will evaluate the pictographs without influence from the primes. They will judge the pictographs on the basis of the primes only to the extent that the prime activates some evaluation and they are unable to control that influence on their judgments. By arranging the task in this way, any misattributions that persist despite the intended task requirements provide evidence of automatic responses to the primes.

In a series of validation studies, misattributions provided valid estimates of attitudes (Payne, Cheng, et al., 2005). The AMP is notable in that it shows high reliability, and in those conditions in which high implicit–explicit correlations were theoretically expected, high correlations have been found. Moreover, in conditions in which implicit and explicit tests were expected to diverge, the AMP showed clear dissociations. These properties make the procedure well-suited for studying relationships between implicit and explicit evaluations. Using an implicit test with a metric that is an evaluation provides a basis for equating many structural differences.

In Studies 1 and 2, we showed that the implicit–explicit correlation increased as structural fit increased. In Study 3, we used a multitrait, multimethod approach to rule out the possibility that high correlations produced by structural fit were artificially inflated by common method variance. Finally, in Study 4, we found that implicit and explicit tests with high structural fit still showed theoretically predicted dissociations, ruling out the idea that high structural fit renders the tests redundant.

Study 1

Our goal in the first study was to compare performance on implicit and explicit tests that varied in structural fit. We expected

that the measures with the greatest structural fit would show the highest implicit–explicit correlations. We surveyed participants using two commonly used explicit measures, the Modern Racism Scale (MRS; McConahay, 1983) and the ATB (Brigham, 1993). We also administered the AMP to measure implicit racial evaluations. The primes for the task consisted of photos of White and Black persons' faces. Participants were shown face primes, followed by Chinese pictographs that they were asked to rate for pleasantness. Unlike the method used in previous studies, participants made their ratings on a continuous 4-point scale from *very unpleasant* to *very pleasant*. This change provided a basis for structural fit with the explicit measures, as will be described in more detail below. As in previous research using the AMP, participants were instructed to evaluate the Chinese pictograph and to avoid being influenced by the primes.

Equating Implicit and Explicit Test Structures

The design described so far is similar to many previous studies assessing implicit–explicit correlations. Although the response metric for the implicit and explicit tasks is the same, the implicit and explicit tasks differ in the kinds of stimuli presented and the processes in which participants must engage. That is, in the questionnaires, participants were asked to endorse or reject complex verbal propositions about racial groups. In the AMP, participants were asked to evaluate the pleasantness of Chinese pictographs after being primed with the faces of White and Black individuals. To equate these features, a second version of the AMP was included. In this version, participants were shown the same prime–target sequences as in the original AMP. However, rather than being told to avoid influence from the prime and evaluate the pictograph, participants were told to avoid the pictograph and evaluate the prime. Figure 1 illustrates the procedure.

We use *indirect evaluation* to refer to the original version of the AMP in which participants evaluate pictographs, because this task provides an indirect measure of reactions toward the primes. In contrast, we use *direct evaluation* to refer to the alternative version, in which participants directly express their evaluations of the primes. The advantage of comparing indirect and direct evaluations in this way is that the tasks are equated on many extraneous factors. In the indirect task, participants intend *not* to display any evaluation of the prime objects. In the direct task, they intend to express evaluations of those same objects. The tasks are equated on the stimuli presented and the type of judgment to be made. They differ only in intent.

By arranging tasks in this way, one can compare the self-report questionnaires with indirect AMP evaluations, which differ both in "implicitness" and in structure. The questionnaires can also be compared with direct AMP evaluations, which differ in structure but are both explicit measures. Finally, we can compare indirect and direct AMP evaluations, which differ in implicitness but are structurally matched. The structural-fit hypothesis predicts that measures with the most similar structures will show the greatest correlations.

Method

Participants

Participants were seventy-five undergraduates (62 women and 13 men) who participated for partial course credit. They ranged in age from 17–21 years (M = 18.46, SD = 0.72). Ethnic groups included 72% White, 17% African American, 2.5% Asian, 7% Hispanic, and 1.5% Native American.

Procedure

Participants were seated at a computer and asked to complete several measures. Of interest were indirect AMP evaluations, direct AMP evaluations, and racial-attitude questionnaires. Indirect and direct evaluations were completed in a counterbalanced order, followed by the questionnaires. Participants next provided demographic information and were debriefed.

AMP

Indirect evaluations. For the indirect rating trials, participants were presented with one of three kinds of primes: a Black face, a

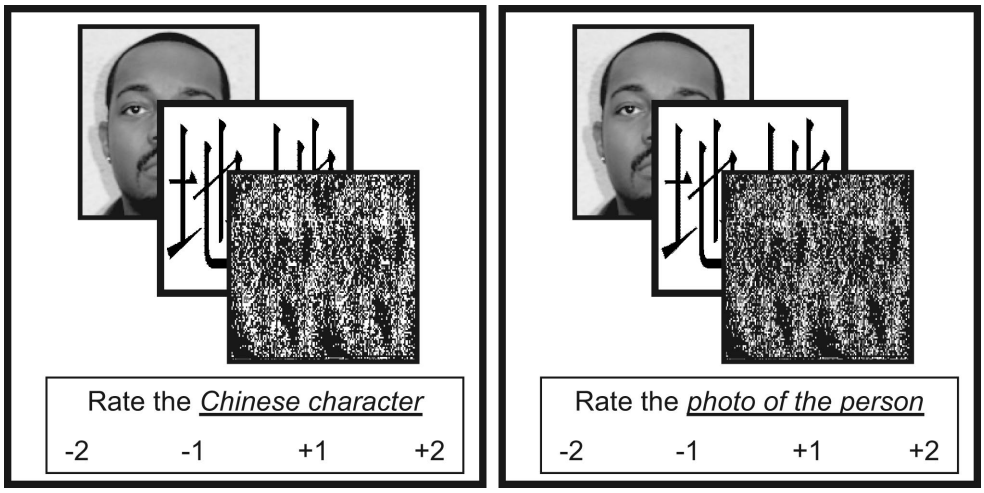

Figure 1. Schematic illustration of affect misattribution procedure (AMP) with indirect and direct ratings.

White face, or a gray square that served as a neutral prime. The face primes were 12 Black men and 12 White men. The pictures showed only the model's face, with a neutral facial expression. Based on pilot testing, the Black and White photos were matched on attractiveness and were selected to be highly prototypical of their respective racial category.

The prime appeared in the center of the screen for 100 ms, followed by a blank screen for 100 ms and then a Chinese pictograph for 100 ms (see Figure 1). Following the pictograph, a patterned mask of black and white "noise" appeared. At the bottom of the screen was a 4-point rating scale that included −2 (*very unpleasant*), −1 (*slightly unpleasant*), +1 (*slightly pleasant*), and + 2 (*very pleasant*). After participants provided their evaluation of the pictograph, the next trial began. A total of 72 randomly ordered trials were presented, with 24 neutral, 24 Black, and 24 White primes paired with 72 unique Chinese pictographs. For each participant, the computer paired a pictograph with a prime in a new random order.

Participants were told that the task was about making judgments while avoiding distraction. They were instructed to rate the pleasantness of the Chinese pictographs using the rating scale. Participants were warned to not let their rating of the pictographs be influenced by the preceding photo. This warning was included to ensure that AMP responses represented the effect of the prime, despite participants' attempts at correction, thereby serving as an indication of the automatic influence of prime-invoked attitudes (Payne, Cheng, et al., 2005). The instructions read as follows:

> For this round of judgments you should ***rate the Chinese characters***. Please note that sometimes the photos flashed prior to the characters can influence people's ratings of the Chinese characters. Please try your best not to be influenced by the photographs. Instead, ***please give us an honest judgment of how pleasant or unpleasant is your reaction to each Chinese character***. Of course, there are no right or wrong answers. Just report your "gut reaction." [Emphasis was in the original.]

Direct evaluations. The direct rating procedure was identical to the indirect rating procedure with three exceptions. The first and most important difference was that participants were instructed to rate their evaluations of the prime photographs and to avoid being influenced by the Chinese pictographs. Because the pictographs were ambiguous and randomly paired with the primes, they could not actually exert any systematic influence on ratings of the primes. The instructions read as follows:

> For this round of judgments you should **rate the photos of people**. Please note that sometimes the Chinese characters flashed after the photos can influence people's ratings of the photos. Please try your best not to be influenced by the characters. Instead, **please give us an honest judgment of how pleasant or unpleasant is your reaction to each person's photo**. Of course, there are no right or wrong answers. Just report your "gut reaction."

The second difference was that no neutral primes were included, because direct evaluations of a gray square would be uninformative. The third difference was that only 24 trials were included, one trial for each unique prime photo. In the direct rating blocks, each prime photo was rated only once for the same reason that each item is only presented once on questionnaires. Because participants were directly expressing their attitudes toward the attitude objects, there was little need for repetitive judgments of the same items.

Self-Report Attitude Measures

Two self-report measures of racial attitudes were used: the MRS (McConahay, 1983) and the ATB (Brigham, 1993). The MRS is a 7-item assessment of anti-Black attitudes and includes items such as "Over the past few years, the government and news media have shown more respect to blacks then they deserve." Responses were made on a 9-point scale ranging from 1 (*strongly disagree*) to 9 (*strongly agree*). The ATB is a 20-item assessment that includes items such as "Black and White people are inherently equal" and "It is likely that Blacks will bring violence to neighborhoods when they move in." Responses were made on a 9-point scale ranging from 1 (*strongly disagree*) to 9 (*strongly agree*).

Results

The key questions concerned whether the correlations between implicit and explicit measures depended on structural fit between measures. But before examining those correlations, we report mean performance on the indirect and direct evaluations. The order of tests did not produce any main effects or interactions and so will not be discussed in the following analyses.

Indirect Evaluations

Pleasantness ratings were averaged for Black primes, White primes, and Neutral primes. Before analysis, we recoded the responses from a −2 to +2 scale to a 1 to 4 scale to simplify analysis and presentation. We analyzed ratings using a repeated-measures analysis of variance (ANOVA). This analysis showed a significant effect of prime on pleasantness ratings, $F(2, 148) = 9.30$, $p < .01$. Ratings were highest for neutral primes ($M = 2.81$, $SD = 0.40$), followed by White primes, ($M = 2.70$, $SD = 0.31$), and they were lowest for Black primes ($M = 2.58$, $SD = 0.40$). The contrast between Black and White primes was significant, $F(1, 74) = 5.21$, $p < .05$, as was each of the contrasts between neutral primes and both Black and White primes, Fs > 4.57, ps $< .05$. These analyses show more positive evaluations of the White primes than the Black primes on the indirect test.

Direct Evaluations

We analyzed pleasantness ratings as above to compare direct evaluations of the Black and White faces. White faces were evaluated significantly more positively ($M = 2.67$, $SD = 0.35$) than Black faces ($M = 2.34$, $SD = 0.40$), $F(1, 74) = 25.44$, $p < .01$. Both direct and indirect measures showed similar preferences for White faces over Black faces at the mean level. But the main questions concerned how these evaluations related to each other and to the questionnaire measures of racial attitudes.

Individual Differences

Scoring. We computed a single score for each person's indirect evaluations by taking the difference between ratings on White-prime trials and ratings on Black-prime trials. Direct evaluations were scored in the same way. Higher scores reflected greater preference for White faces relative to Black faces. We scored the ATB and MRS scales by taking the mean of responses after reverse coding, where appropriate. The mean score for the ATB

was 7.00 (SD = 1.22). The mean score for the MRS was 2.63 (SD = 1.25). Higher scores on the ATB reflect more positive attitudes toward Black people, whereas higher scores on the MRS reflect more negative attitudes toward Black people. For the purpose of the following analyses, the ATB scale was reverse-scored so that all measures were scored in the same direction, with higher scores reflecting more negative attitudes toward Black people.

We calculated reliability for indirect ratings by taking a difference score between pleasantness ratings on each White-prime trial and the rating on a randomly paired Black-prime trial. This produced 24 difference scores that were treated as items in a reliability analysis (for a fuller description, see Payne, Cheng, et al., 2005). Reliability for direct ratings was computed in the same way. Reliability was acceptable for all measures (indirect α = .69; direct α = .71; ATB α = .80; MRS α = .90).

Correlations. We hypothesized that over and above the implicit–explicit distinction, the attitude tests with greatest structural fit would show the strongest correlations. The two tests with the highest degree of fit were the two explicit questionnaires (ATB and MRS). These questionnaires were matched in both structural fit (i.e., they asked similar kinds of questions) and in the fact that both were explicit measures. Table 1 displays the correlations among all tests. As expected, the questionnaires were strongly correlated with each other. In contrast, the ATB and MRS were only weakly, though significantly, correlated with indirect evaluations. These findings are consistent with much previous research that has showed small implicit–explicit correlations for racial attitudes. The traditional way to interpret this finding is that implicit and explicit tests reflected different attitudes or different processes. That is, the reason for implicit–explicit divergence is said to lie in the difference between implicitness and explicitness. However, the questionnaires differed from the indirect ratings not only in implicitness, but also in structure (e.g., pictures vs. verbal statements, exemplars vs. groups, simple pleasantness judgments vs. opinions on broad policies, etc.). Because they shared neither implicit–explicitness nor structural features, the low correlations may result from either kind of difference.

The traditional interpretation would predict that direct evaluations should be highly correlated with the ATB and MRS because they are all explicit measures. However, as shown in Table 1, direct ratings were no more strongly correlated with these questionnaires than the indirect ratings were. We suggest these low correlations reflect the lack of structural fit between the direct ratings and the questionnaires. If this hypothesis is correct, then the direct and indirect tests should be strongly correlated, because despite differing in implicitness, they are equated in structure. In fact, the correlation between direct and indirect tests (r = .64) was nearly as large as the correlation between the ATB and MRS scales (r = .68), which were equated on both dimensions. These two correlations were not significantly different from each other. They were both, however, significantly greater than the other four correlations (all ps < .05). This pattern of correlations supports the proposal that structural fit is an important factor influencing the size of the correlation between implicit and explicit tests.

Table 1
Correlations Among Self-Report Scales of Racial Attitudes and Direct and Indirect AMP Ratings, Study 1

Measure	ATB	MRS	Direct AMP	Indirect AMP
ATB	—	.68*** (.54–.79)	.25* (.02–.46)	.25* (.02–.46)
MRS		—	.26* (.04–.46)	.24* (.01–.45)
Direct AMP			—	.64*** (.49–.77)

Note. In parentheses are 95% confidence intervals for each correlation. AMP = affect misattribution procedure; ATB = Attitudes Toward Blacks Scale; MRS = Modern Racism Scale.
* $p < .05$. *** $p < .001$.

Discussion

Some aspects of these results replicated commonly observed findings. The two explicit questionnaires were highly correlated with each other. They were also quite weakly correlated with an indirect test of race attitudes. Both of these facts can be explained by two accounts. The traditional account is that the two questionnaires were strongly related because both tests explicitly asked participants to report their attitudes. In contrast, the questionnaires were not strongly related to the indirect test because the questionnaires are explicit measures, whereas the indirect test is an implicit measure.

The structural-fit account can also explain these findings. By that account, the questionnaires were strongly related to each other because they were well-matched on measurement features. For example, they both asked participants to endorse or reject propositions about Black people as a group in American society. In contrast, the questionnaires were weakly related to the indirect test because they differed in measurement features. Unlike the questionnaires, the indirect test asked participants to make simple feeling-based judgments of pleasantness. And the primes were faces of individuals, not social-group labels. So both a traditional account based on implicitness and a structural-fit account can explain the strong correlation between questionnaires and the weak correlation between the questionnaires and the indirect rating.

We must account for two other cells in our design. First is the correlation between the questionnaires and the direct test. If the implicit–explicit distinction were the only factor at work, then we would expect this correlation to be high because both are explicit tests. However, these weak correlations are more consistent with a structural-fit explanation. Although these tests are all explicit measures of race attitudes, they differ in their measurement features in much the same ways that the questionnaires differ from the indirect test. That is, rating feelings toward pictures of individuals is quite different from expressing attitudes toward social policies regarding racial groups. Finally, the strong correlation between indirect and direct tests presents a puzzle for any account based only on the implicit–explicit distinction. This correlation is, however, consistent with the structural-fit account. These results suggest that when implicit and explicit tests are equated on extraneous measurement features, they may be much more highly correlated than previously thought.

Some Notes on Scaling

Despite using the same response scale, indirect and direct AMP ratings tend to differ in extremity. One reason is that on indirect

ratings, participants rate Chinese pictographs, which are selected to be fairly neutral. As a result, indirect ratings will tend toward neutral values at the mean level. Participants are also instructed to avoid any influence of the primes. Although participants cannot do so completely (Payne, Cheng, et al., 2005), partial success will tend to reduce the magnitude of differences between Black- and White-prime trials. These features make the task a conservative test of racial bias and also increase confidence that any race bias that persists is beyond voluntary control. At the same time, these considerations mean that a 1.5 indirect rating (of a pictograph) does not necessarily mean the same as a 1.5 direct rating (of a human face). For this reason, it would be inappropriate to directly compare the mean levels across the two tasks. The value of making the task structures more matched is not found in simply comparing raw numbers across tasks. Instead, the value is in encouraging similar cognitive processes on the two tasks, except for the systematic differences of interest. Individual-difference correlations are more informative for testing our hypotheses than are the absolute mean levels. (The two tasks can, of course, be standardized, which removes any meaningful mean difference and does not change the correlations). The correlations observed here confirm that the individual differences were systematic and interpretable.

Broadening the Conclusion

In Study 1, we took a first step toward exploring the importance of structural correspondence. But there are some idiosyncratic aspects of the measures that need to be broadened. For example, all of the explicit measures were verbal, whereas the implicit measure was based on pictures. This difference was chosen as a way to manipulate structural correspondence, but it is important to show that the strong implicit–explicit correlation observed is not limited to pictorial methods. Our structural-fit analysis leads us to predict that when verbal group labels such as "Black" and "White" are used as primes, both indirect and direct ratings should be correlated with other measures that ask participants to evaluate similar verbal categories. In Study 2, we aimed to replicate these findings and extend them by manipulating structural fit across a wider range of items to be evaluated, including group labels.

Study 2

Study 2 was designed to incorporate one of the most popular ways to measure attitudes: the "feeling thermometer." In this simple method, participants are asked to rate their feelings toward, say, African Americans, on a scale ranging from *very cold and unfavorable* to *very warm and favorable*. If the same verbal labels and rating scale are used for direct ratings using the AMP, we have essentially re-created a feeling thermometer. And if the same labels and scales are used for indirect AMP ratings, we have an *implicit feeling thermometer*. The two measures can be matched on all the relevant features as already described, and they differ only in the intent to express an evaluation of the primes. In this study, we manipulated structural fit by comparing evaluations on the basis of (a) faces, (b) group labels, (c) the ATB and MRS scales, and (d) a traditional feeling thermometer. Our hypothesis was that more closely matched test structures would reveal larger implicit–explicit correlations.

Method

Participants

Participants were forty-eight undergraduates (28 women and 20 men) who participated for partial course credit. They ranged in age from 17–21 years ($M = 18.71$, $SD = 0.87$). Ethnic groups included 77% Caucasian, 15% African American, 6% Asian, and 2% Hispanic.

Design and Procedure

Participants completed the AMP in four blocks under the same instructions as described in Study 1. The four blocks (indirect/pictures, indirect/group labels, direct/pictures, direct/group labels) were counterbalanced for order (no order effects emerged). On indirect blocks, participants were warned that the primes might influence their judgments and that they should try their best to avoid any such influence. On the direct blocks, they were warned that the pictographs might influence their judgments and that they should try their best to avoid that influence. After the AMP procedure, participants completed self-report measures of racial attitudes, provided demographic information, and were debriefed.

AMP

Structure: Face versus word primes. For some blocks of trials, participants were shown photographs of Black and White young men prior to the pictographs (neutral primes were not used in this study, because they were not informative for individual differences). The photographs were the same as those used in Study 1. For the other blocks, participants were primed with words rather than photographs. Specifically, they were shown three White group labels (European Americans, White Americans, and Whites) and three Black group labels (African Americans, Black Americans, and Blacks), one at a time, followed by a pictograph.

Direct versus indirect evaluations. For the indirect trials, participants evaluated the Chinese pictographs, ignoring their feelings toward the primes. There were 48 indirect trials for the face primes and 48 indirect trials for the word primes. For the direct trials, participants rated the primes instead. Thus, in the face-prime condition, participants rated their feelings toward the photographs, ignoring their feelings toward the pictographs. Participants rated each photo once, for a total of 24 direct/picture trials. In the word-prime condition, participants rated their feelings toward the verbal group labels, ignoring the pictographs. They rated each group label once, for a total of six direct/verbal trials.

To ensure that participants could follow instructions without becoming confused about what they were supposed to rate, we presented a prompt on every trial that reminded them what to evaluate. On indirect blocks, the instruction "Rate feelings toward Chinese character" appeared just above the 4-point rating scale on each trial. On direct blocks, the phrase "Rate feelings toward photo of person" or "Rate feelings toward social group" appeared.

Self-Report Racial-Attitude Measures

Participants completed the MRS (McConahay, 1983) and the ATB (Brigham, 1993). A traditional feeling thermometer scale was also used, in which participants rated their feelings toward

four racial groups, including White Americans, Asian Americans, Black Americans, and Hispanic Americans. Ratings were made on a 9-point scale ranging from 1 (*very cold and unfavorable*) to 9 (*very warm and favorable*). Feelings toward Black people and White people were of primary interest.

Operational Definition of Structural Fit

Structural fit in this study was manipulated by the similarity of the attitude objects to be evaluated. For the purpose of analyzing how implicit–explicit correlations change depending on structural fit, rank orders were assigned (a priori) to the similarity between each implicit–explicit pair. The rank orders were different for the two implicit tests because of their different structures. For the indirect/picture task, the most similar explicit task was the direct/picture task (rank = 1), followed by the direct/group-label task (rank = 2), and thermometer (rank = 3), and both ATB and MRS were assigned a tied ranking (rank = 4). The latter two were assigned a tie, because there is little a priori reason for distinguishing between them in similarity to the AMP tasks. Task similarity ranged, then, from pictures of group members to verbal labels for the groups to more abstract policy-focused scales. For the indirect/group-labels task, the most similar explicit task was the direct/group-labels task (rank = 1), followed by the thermometer (rank = 2), both ATB and MRS scales (tied at rank = 3), and the direct/picture task (rank = 4). Task similarity in this case ranged from highly similar group labels (in the direct/group-labels task and traditional feeling thermometer) to policy-focused verbal questions concerning entire groups to pictures of individual group members. The rank ordering of similarity provided a way to test whether the implicit–explicit correlation depended on structural fit. The question was not whether any particular pair of correlations differed from each other, but instead whether implicit–explicit correlations showed a general trend to increase as structural fit increased.

Results

Because our main hypotheses concerned individual difference correlations, we summarize the mean ratings briefly and then focus in more depth on individual differences. As in Study 1, ratings were recoded to a 1–4 scale.

Indirect Evaluations

This sample did not show a significant mean difference in their indirect responses to Black versus White primes, nor for pictures, nor for verbal labels. There was no significant difference between ratings of Chinese pictographs when primed with Black faces ($M = 2.58$, $SD = 0.41$) versus White faces ($M = 2.55$, $SD = 0.34$), $F(1, 47) = .32$, $p = .58$. Nor was there a significant difference when primed with Black verbal labels ($M = 2.75$, $SD = 0.53$) versus White verbal labels, ($M = 2.66$, $SD = 0.44$) $F(1, 47) = 1.06$, $p = .31$.

Direct Evaluations

A similar pattern emerged for direct ratings, with ratings of Black faces ($M = 2.35$, $SD = 0.46$) slightly but nonsignificantly lower than ratings of White faces, ($M = 2.51$, $SD = 0.44$), $F(1, 47) = 3.79$, $p = .06$. Direct ratings of the group labels were similar for Blacks ($M = 3.01$, $SD = 0.68$) and Whites ($M = 3.16$, $SD = 0.63$), $F(1, 47) = 1.05$, $p = .31$. Finally, the traditional feeling thermometer showed no difference between feelings toward Blacks ($M = 6.50$, $SD = 1.73$) versus Whites ($M = 6.70$, $SD = 1.86$), $F(1, 47) = .64$, $p = .43$. At the mean level, there was little or no race bias on indirect ratings and direct ratings, nor on a traditional feeling thermometer. The difference in mean levels of bias compared with Study 1 is likely a consequence of sampling error, combined with the smaller sample size in Study 2. Completing multiple race-related tasks may also have reduced bias by making race a salient topic or via practice effects. Nonetheless, mean levels are not of interest for our theory-driven hypotheses concerning individual differences, to which we turn next.

Individual Differences

Implicit–explicit correlations. Individual scores were computed for each measure as described in Study 1. All measures showed good reliability as estimated with Cronbach's alpha (ATB = .92, MRS = .90, indirect/verbal = .84, indirect/picture = .71, direct/verbal = .80, direct/picture = .74). Even though this sample evaluated Whites and Blacks about equally on average, the reliability estimates suggest that individual differences were systematic and reliable.

We first consider implicit–explicit correlations concerning the indirect/picture task, shown in Table 2. Rather than reporting all possible pairwise comparisons between correlation coefficients, we report the 95% confidence interval for each correlation in Table 2 so that differences between any pair of correlations can be inspected. The correlations ranged from .21 to .48. As shown by the confidence intervals, the highest correlation was significantly different from the lowest correlation, but the intermediate correlations were not significantly different from each other. A similar pattern emerged for implicit–explicit correlations concerning the indirect/group-labels task, shown in Table 3. The most similar tasks showed a correlation of .65, in contrast to .39 for the least similar pair. The highest coefficient was significantly different from the lowest two coefficients, as shown by the confidence intervals.

The main question in this study was not about the difference between any specific pair of correlations but, rather, the larger

Table 2
Implicit–Explicit Correlations Between the Indirect/Picture Test and Each Explicit Test, in Rank Order of Structural Fit, Study 2

Fit (rank)	Test	Indirect/picture
1	Direct/picture	.48*** (.23–.68)
2	Direct/group labels	.39** (.12–.61)
3	Thermometer	.42** (.16–.64)
4	ATB	.35* (.07–.58)
4	MRS	.21 (−.08–.47)

Note. In parentheses are 95% confidence intervals for each correlation. Thermometer = feeling thermometer; ATB = Attitudes Toward Blacks Scale; MRS = Modern Racism Scale.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Implicit–Explicit Correlations Between the Indirect/Labels Test and Each Explicit Test, in Rank Order of Structural Fit, Study 2

Fit (rank)	Test	Indirect/group labels
1	Direct/group labels	.65*** (.45–.79)
2	Thermometer	.48*** (.23–.68)
3	ATB	.46*** (.21–.67)
3	MRS	.38** (.11–.60)
4	Direct/pictures	.39** (.12–.61)

Note. In parentheses are 95% confidence intervals for each correlation. Thermometer = feeling thermometer; ATB = Attitudes Toward Blacks Scale; MRS = Modern Racism Scale.
** $p < .01$. *** $p < .001$.

trend: Do implicit–explicit correlations increase as structural fit increases? To answer this question, we reverse scored the rank orders so that higher values represent greater structural fit. We then plotted the size of the implicit–explicit correlation (including both indirect tests) against the degree of structural fit. The results are shown in Figure 2. Each point on this scatter plot represents one of the implicit–explicit correlations in Tables 2 and 3. We tested the correlation between structural fit and the size of implicit–explicit correlations using Spearman's rank-order coefficient, which showed a very strong relationship, r_s (10) = .90, $p < .001$. The degree of relationship between implicit and explicit tests was tightly linked to their structural fit.

Other correlations. The correlation between the two indirect AMP tests was $r = .50$, $p < .001$. This correlation is higher than is often seen when comparing different implicit measures, but it is similar to that reported when latent variable analysis was used to correct for random measurement error (Cunningham et al., 2001). Table 4 shows the correlations among explicit tests. Although the effects of structural fit theoretically apply to explicit–explicit correlations also, it is difficult to know how to clearly rank their structural similarities. These tests were selected so that they clearly varied in similarity to the implicit tests, but they were not selected to be clearly ranked in similarity to each other. In general, measures that were highly similar, such as ATB and MRS, were highly correlated. The correlation between the thermometer and the direct/group-labels task was also high. The direct/picture task tended to have lower correlations with the verbal measures, which were less similar. Because it is difficult to rank many of the other pairs, however, we remain cautious about drawing conclusions about interrelations among the explicit tests.

Discussion

In Study 2, we tested the hypothesis that implicit and explicit measures of racial attitudes can be highly correlated when they are equated on structural features. The study, in which we used several measurement techniques, supported the hypothesis. The relation between implicit and explicit measures steadily increased as structural similarity increased. Phrased another way, comparing implicit and explicit measures that differed in irrelevant features undermined their correlation. Had we looked only at an implicit measure and an explicit measure that differed on many structural features (as is typically done), we would have wrongly concluded that the underlying attitudes were only weakly related.

The manipulation of the items to be evaluated is, of course, only one of many possible ways that tests may differ. Commonly used implicit and explicit tests often differ in several ways at once. The correlation between implicit and explicit tests is important because it has been a key piece of evidence for theories about the nature of implicit evaluation. The current findings shed new light on that evidence by suggesting that comparisons on the basis of these tests may severely underestimate the relationship between underlying implicit and explicit evaluations.

Our argument that poor structural fit underestimates implicit–explicit correlations can be seen as an instance of unshared method variance. On the one hand, unshared method variance can cause true correlations to be underestimated (Bagozzi & Yi, 1991; Campbell & Fiske, 1959; Podsakoff et al., 2003). Our proposal to equate tests on structural features can help reduce that problem. On the other hand, creating tests with structural fit also increases shared method variance. Shared method variance, in turn, can potentially inflate correlations between measures. Are well-equated test structures a cause for concern?

There is reason to think that the risks of underestimating the implicit–explicit correlation outweigh the risks of overestimating it. One reason is that comparing implicit and explicit tests can be thought of just as any other within-subjects experimental design. The ability to draw conclusions about a manipulated variable rests on the assumption that other variables do not also differ between conditions. Holding such extraneous factors constant across experimental conditions is rarely considered a threat to validity in other experimental research. Instead, it represents good experimental control.

A second reason is that the methods used here control for some of the common sources of method variance that may inflate correlations. One common source of method variance is a general response bias. For example, some participants in our studies may simply like Chinese pictographs more than others. Or some people may simply use the higher range of a response scale, whereas

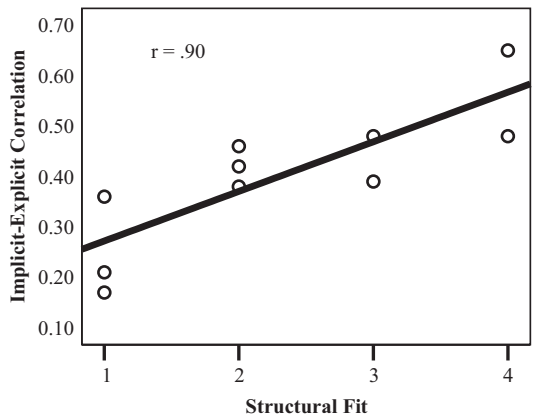

Figure 2. The size of the implicit–explicit correlation as a function of structural fit (reverse rank order) in Study 2.

Table 4
Correlations Among Explicit Tests, Study 2

Test	MRS	Thermometer	Direct/group labels	Direct/picture
ATB	.81*** (.68–.89)	.66*** (.46–.80)	.59*** (.38–.76)	.39** (.12–.61)
MRS	—	.54*** (.31–.72)	.49*** (.24–.68)	.25 (−.04–.50)
Thermometer		—	.72*** (.56–.83)	.50*** (.26–.69)
Direct/group labels			—	.39** (.12–.61)

Note. In parentheses are 95% confidence intervals for each correlation. MRS = Modern Racism Scale; Thermometer = feeling thermometer; ATB = Attitudes Toward Blacks Scale.
** $p < .01$. *** $p < .001$.

others use the lower range. The scoring algorithm we have used in this article, however, controls for general response biases by comparing Black-prime trials with White-prime trials using a difference score. This approach avoids some sources of method bias in the same way that including control groups in experimental designs avoids systematic differences between conditions that are unrelated to the experimental manipulation.

Of course, when methods in two conditions are matched, there are a potentially infinite number of ways that they could be similar. There is no way on conceptual grounds to rule out all possible sources of shared method variance that could inflate their relationships. So there remains a risk that structural fit may overestimate correlations. Therefore, we developed for Study 3 a multitrait, multimethod design to test whether these relationships are heavily influenced by shared method variance in practice.

Study 3

For this study, we selected a second attitude object that we expected to be unrelated to race attitudes. Previous research in our lab showed that implicit and explicit attitudes toward alcohol measured using the AMP showed correlations comparable to the correlations reported here for race attitudes (Payne, Govorun, & Arbuckle, in press). That is, implicit and explicit alcohol attitudes showed moderate correlations overall, which increased when social pressure encouraged honest reporting but decreased when social pressure encouraged underreporting drinking. Attitudes toward alcohol, however, seemed unlikely to be related to racial attitudes. Therefore, in this study we measured attitudes toward beer and race using direct and indirect AMP ratings. We expected direct and indirect tests of racial attitudes to correlate as in the previous studies. We also expected direct and indirect tests of alcohol attitudes to correlate with each other. If shared method variance artificially inflates the relationships between direct and indirect ratings, then the similar methods should produce large correlations between alcohol and race attitudes. But if not, then there should be evidence of discriminant validity.

Method

Participants

Participants were 66 undergraduates (46 women and 20 men) who participated for partial course credit. They ranged in age from 18–22 years ($M = 18.76$, $SD = 0.98$). Ethnic groups included 64% White, 20% African American, 10% Hispanic, and 6% Asian. Three participants were excluded because of their ability to recognize Chinese characters.

Procedure

Participants completed the AMP in four blocks under the same instructions as Study 1. We presented the four blocks (indirect/race, indirect/drinks, direct/race, and direct/drinks) in a random order to prevent systematic order effects. On indirect blocks, participants were warned that the primes might influence their judgments and were instructed to try their best to avoid any such influence. On direct blocks, they were warned that the pictographs might influence their judgments of the primes and were instructed to try their best to avoid that influence.

AMP

Race and drink primes. For some blocks of trials, participants were shown photographs of Black and White young men prior to the pictographs (as in Study 2, neutral primes were omitted). The photographs were the same as those used in Studies 1 and 2. For the other blocks, participants were primed with 12 photographs of beer and 12 photographs of drinking water (no neutral primes were used). These photographs were taken from a previous study (Payne et al., in press), in which they had been matched for visual attractiveness.

Direct versus indirect evaluations. For the indirect trials, participants were instructed to evaluate the Chinese pictographs, ignoring their feelings toward the prime. There were 72 indirect trials for the race primes and 72 indirect trials for the drink primes. For the direct trials, participants were instructed to rate the prime. Thus, in the race-prime blocks, participants rated their feelings toward the photographs of Black and White men, ignoring their feelings toward the pictographs. Participants rated each photo once, for a total of 24 direct/race trials. In the drink-prime blocks, participants rated their feelings toward the photographs of beer and water, ignoring the pictographs. They rated each photo once, for a total of 24 direct/drink trials. To avoid confusion about what participants were supposed to rate, we gave them new instructions before each block. In addition, a prompt was presented on each trial that reminded them what item to evaluate, as in Study 2.

Results

Mean Results

At the mean level, indirect ratings showed a nonsignificant tendency toward preference for water over beer, ($M = 0.28$, $SD = 0.65$ vs. $M = 0.08$, $SD = 0.71$), $F(1, 62) = 2.55$, $p = .12$. There was also a nonsignificant mean preference on White- versus Black-prime trials ($M = 0.10$, $SD = 0.57$ vs. $M = 0.07$, $SD = 0.63$), $F(1, 62) = .13$, $p = .72$. Direct ratings showed a nonsignificant preference for water over beer, ($M = 0.54$, $SD = 0.99$ vs. $M = 0.17$, $SD = 1.21$), $F(1, 62) = 2.45$, $p = .12$, and a significant preference for White faces over Black faces ($M = -0.13$, $SD = 0.68$ vs. $M = -0.38$, $SD = 0.72$), $F(1, 62) = 4.15$, $p = .05$. Direct and indirect attitude scores were computed for each individual by subtracting responses on Black trials from those on White trials and subtracting responses on beer trials from those on water trials so that higher scores indicate greater relative preferences for White faces and water, respectively.

Convergent and Discriminant Validity

Before examining the individual difference correlations, we removed three outlier scores because they were 3.5 standard deviations or more from their respective means. Table 5 displays the correlation matrix, with the reliabilities of each test displayed on the diagonal. As expected, direct and indirect measures of race attitudes were significantly correlated, as were direct and indirect measures of drink attitudes. However, there were no significant correlations between direct ratings of race attitudes and direct ratings of drink attitudes, and there were no significant correlations between indirect ratings for the two attitude topics. Finally, there were no significant correlations between indirect ratings of one attitude object and direct ratings of the other attitude object.

To more formally establish convergent and discriminant validity, we performed an exploratory factor analysis with oblique rotation. This analysis converged on a two-factor solution, with race attitudes loading highly on one factor and drink attitudes loading highly on the other factor (see Table 6). These two factors were uncorrelated, $r = -.05$. The pattern of correlations seems to provide clear evidence that correlations were not created simply by using similar methods.

Table 5
Correlations Among Indirect and Direct Tests of Attitudes Toward Race and Drinks

	Race		Drink	
Variable	Indirect	Direct	Indirect	Direct
Race				
Indirect	(0.75)			
Direct	.44**	(.78)		
Drink				
Indirect	.03	.03	(.90)	
Direct	−.09	−.20	.41**	(.95)

Note. Values in parentheses are reliabilities for each test.
** $p < .01$.

Discussion

In Study 3, we found evidence for convergent and discriminant validity. If the AMP inflated correlations simply because of shared method variance, then the correlations between beer and race attitudes measured with the same method should have been substantial. The results, however, showed no evidence for this account. This helps rule out the possibility that structurally equated measures correlate with each other spuriously because of shared method variance. Instead, this study showed that the measures cohered on the basis of the attitude object, not the method of measurement. The evidence suggests that structural fit removes barriers to correlations, but we found no evidence that it artificially inflates them.

Although Study 3 addressed the potential concern of shared method variance, there is another potential concern that should be addressed. The concern is that structurally matched tests might not successfully differentiate implicit or automatic responses from explicit or controlled responses. On the one hand, it could be that the indirect ratings were not truly implicit. Perhaps participants disregarded the warning to avoid evaluating the primes and intentionally responded on the basis of their evaluations of the primes. If that is the case, then the correlations reported do not reflect implicit–explicit relationships at all. On the other hand, it could be argued that the direct ratings were not fully explicit. Perhaps because the primes were flashed briefly, the direct ratings captured more automatic responses than is typical with self-report measures. If that is the case, then the direct picture ratings might be better regarded as a second implicit test.

We conducted a fourth study to rule out these alternative explanations by showing that direct ratings perform as we expect explicit attitude tests to perform, and indirect ratings perform as we expect implicit tests to perform. To do so, we selected a variable known to differentially affect explicit versus implicit tests: social pressure to avoid showing prejudice. If direct ratings operate as an explicit measure and indirect ratings act as an implicit measure, they should respond differently to such a manipulation. The vulnerability of explicit tests—and the resistance of implicit tests—to such pressure is one of the most common reasons for using an implicit test. In Study 4, we manipulated the amount of social pressure, with the expectation that it would affect direct ratings but not indirect ratings.

Study 4

For this study, we used only direct and indirect AMP ratings with faces as primes. During indirect ratings, participants evaluated pictographs while trying to avoid influence from the primes. For direct ratings, they rated the face primes themselves. The same prime and target stimuli were presented for both kinds of trials, and judgments were provided on identical scales. The only difference was that on direct ratings, participants intended to express their evaluations of the faces, whereas on indirect ratings, they intended not to. Because they had good structural fit, we expected these measures to be significantly related. But the key to this study was that one group of participants was encouraged to respond honestly and to ignore any social pressure to avoid bias. The other group was encouraged to be vigilant against the possibility of showing subtle race biases. We expected this manipulation to selectively

Table 6
Factor Analysis Loadings in Study 3

Variable	Factor 1	Factor 2
Race		
Indirect	.79	−.14
Direct	.86	.16
Drink		
Indirect	.27	.85
Direct	−.35	.72

affect direct ratings but not indirect ratings. As a result, we predicted a strong implicit–explicit correlation only in the low-pressure condition, where participants could feel free to overtly express their attitudes.

Method

Participants

Participants were 71 undergraduates (53 women and 18 men) who participated for partial course credit. Ethnic groups included 75% Caucasian, 7% African American, 11% Asian, 4% Hispanic, and 3% Native American.

Design and Procedure

The experiment was a 2 (High vs. Low social pressure) × 2 (Direct vs. Indirect rating) design, with social pressure manipulated between participants. Participants completed the direct and indirect AMP ratings under the same instructions described for Study 1. Direct and indirect ratings were completed in separate blocks, counterbalanced for order. For the indirect ratings, a total of 72 randomly ordered trials was presented, with 24 neutral, 24 Black, and 24 White primes. The neutral prime was a gray square, whereas the face primes included 12 White faces and 12 Black faces, each presented twice. Seventy-two unique pictographs were presented and were evaluated on a 4-point scale ranging from −2 (*very unpleasant*) to +2 (*very pleasant*). For the direct ratings, each of the same 12 White and 12 Black faces was presented once, paired with a new set of unique pictographs.

Social Pressure Manipulation

Before beginning the AMP, participants were given instructions that manipulated the social pressure to avoid expressing prejudice. Participants were randomly assigned to receive the low-pressure or high-pressure instructions. In the low-pressure condition, participants received instructions emphasizing that everyone's opinion is valid. The instructions were as follows:

> Race relations are a very complex issue, and people vary widely in their opinions. People have many reasons for their views, including their own personal histories and experiences. Because everyone's experiences are unique, it is important to realize that each individual's perspectives should be respected. Your opinions are valuable to us as they are. We ask that you express your own attitudes and opinions as honestly as possible, even if they are not "politically correct." Remember, your responses will be kept completely confidential—neither your name nor other identifying information will be attached to the data entered into the computer.

In the high-pressure condition, participants were warned to avoid racial bias. The instructions read as follows:

> Race relations are a very important issue, because prejudice and discrimination continue to exist, sometimes in subtle ways. One way that people can overcome the scourge of prejudice is by continually being vigilant for biased tendencies in their own attitudes, opinions, and behavior. Your opinions are important to us. We ask that you express your own attitudes and opinions, keeping in mind the possibility that we are all vulnerable to racial biases.

After completing the AMP phase of the study, participants provided demographic information and were debriefed.

Results

Our hypothesis was that social pressure would affect direct ratings but not indirect ratings, resulting in higher implicit–explicit correlations in the low-pressure condition than in the high-pressure condition. We tested this hypothesis with multiple-regression analyses. First, we gave each participant a score for indirect ratings by taking the difference between indirect ratings on Black- versus White-prime trials. We also gave each participant a score for direct ratings using the same procedure. For both, positive values reflect a preference for White over Black, whereas negative values represent a preference for Black, and zero represents no difference. At the mean level, there was a pro-White bias on both indirect ratings ($M = 0.11$, $SD = 0.36$) and direct ratings ($M = 0.29$, $SD = 0.51$). One-sample t tests showed that the difference from zero was significant for both indirect ratings, $t(70) = 2.59$, $p < .05$, and direct ratings, $t(70) = 4.75$, $p < .01$. Neither indirect nor direct ratings were significantly affected by the social pressure. The indirect test showed a similar preference for Whites over Blacks in the low-pressure condition ($M = 0.07$, $SD = 0.35$) and the high-pressure condition ($M = 0.15$, $SD = 0.38$), $t(69) = .93$, $p = .36$. The direct test also showed a similar mean preference in low-pressure ($M = 0.26$, $SD = 0.59$) and high-pressure conditions ($M = 0.32$, $SD = 0.42$) $t(69) = .51$, $p = .61$. These means should, however, be interpreted in the context of the interaction reported below.

Our hypothesis predicted a significant interaction between indirect ratings and social pressure, because the effect of social pressure should depend on participants' automatically activated attitudes. A majority of the sample (40 participants; 56.3%) showed some degree of pro-White bias on indirect ratings. A sizable minority, however, showed a pro-Black bias (24 participants; 33.8%). Seven participants (9.9%) showed exactly equal indirect ratings on Black and White trials. Participants with automatic pro-White biases were expected to adjust their responses on direct ratings in a pro-Black direction, toward equal ratings of Black and White faces. Participants with pro-Black automatic biases, however, were expected to adjust their responses in a more pro-White direction, again toward equal ratings (the social pressure instructions did not encourage pro-Black responses; they encouraged vigilance against bias of any kind). This pattern would result in a lower correlation between indirect and direct measures in the high-pressure condition, compared with the low-pressure condition.

Figure 3 displays the results. The direct ratings are shown as a function of social-pressure group and indirect ratings. For both direct and indirect ratings, the figure displays raw difference scores, in which higher scores reflect pro-White bias, and a score of zero reflects equal ratings on Black and White trials. In the low-pressure condition, indirect ratings were strongly related to direct ratings, $r(35) = .71$, $p < .001$. In contrast, the relationship was much weaker in the high-pressure condition, $r(36) = .31$, $p = .06$. Direct and indirect ratings thus seem to have measured similar reactions in the low-pressure condition but very different reactions in the high-pressure condition.

We used multiple-regression analysis to test the effect of social pressure on direct ratings in interaction with indirect ratings. The dependent variable was direct ratings, with indirect ratings and social-pressure condition (contrast coded as 1 and −1) entered in the first step and their interaction entered in the second step. The analysis showed a main effect of indirect ratings, indicating that indirect and direct ratings were significantly correlated, $\beta = .53$, $t(68) = 5.10$, $p < .01$. The main effect of social pressure was not significant, $\beta = .003$, $t(68) = .03$, $p = .98$. That is, the social-pressure manipulation did not create an overall tendency to evaluate Black faces more favorably than White faces. Instead, the effect of social pressure to avoid prejudice depended on participants' automatic biases, as shown by a significant Indirect × Social Pressure interaction, $\beta = -.31$, $t(67) = 3.08$, $p < .01$. The inclusion of the interaction term in the second step produced a significant increment in R^2, $F(3, 70) = 13.06$, $p < .001$, for a total R^2 of .37. For participants with an automatic preference for White faces over Black faces (i.e., on the right side of Figure 3), social pressure reduced that bias on direct ratings, bringing them closer to zero. For those with a pro-Black preference (i.e., participants on the left of Figure 3), social pressure also reduced that bias, raising direct ratings closer to the neutral point of zero.

To better clarify the source of these differences, we ran the regression analysis again, analyzing direct ratings of Black faces and White faces separately. The Indirect × Social Pressure interaction was significant for ratings of Black individuals, $\beta = .24$, $t(67) = 2.30$, $p < .05$, but not for ratings of White individuals, $\beta = -.12$, $t(67) = .99$, $p = .33$. Responses to social pressure were thus driven mainly by ratings of Black individuals.

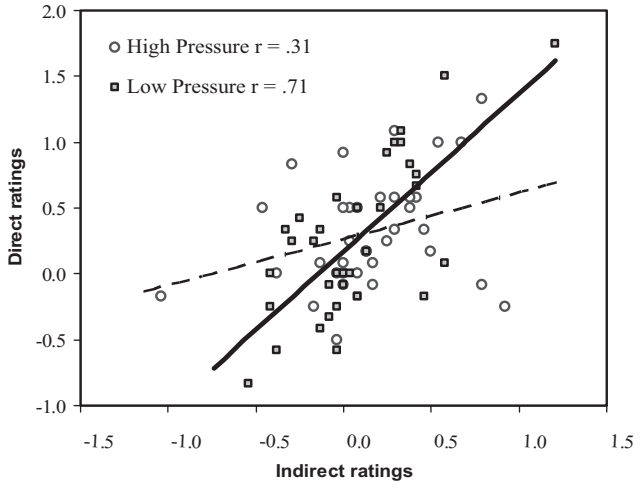

Figure 3. Scatter plot relating indirect affect misattribution procedure (AMP) ratings to direct AMP ratings in high-pressure and low-pressure groups in Study 3.

Discussion

Social pressure had opposing effects, depending on participants' implicit biases. For participants with a pro-White bias, high social pressure led to expression of less bias on the direct ratings than did low pressure. This resulted in direct ratings closer to zero, reflecting equivalent evaluations of Black faces and White faces. In contrast, for those with a pro-Black implicit bias, high pressure led to higher scores on the direct ratings, again resulting in direct ratings closer to the neutral point of zero. In sum, social pressure to avoid prejudice pushed direct ratings toward the neutral point, at which White and Black faces were evaluated similarly.

Whereas Studies 1–3 demonstrated that implicit and explicit tests show substantial correlations when equated in structure, Study 4 ruled out alternative explanations for those findings. There was no evidence that the indirect test could be easily controlled, nor that the direct test could not be controlled. The results of this study provide converging evidence that our methods reflect implicit and explicit evaluations, as they were designed to measure. Social pressure to avoid race bias produced more egalitarian responding on direct ratings of Black and White faces. As a result, direct and indirect ratings were strongly correlated when social pressure was minimized but weakly correlated when social pressure was heightened.

At first blush, it may seem counterintuitive that participants scoring low on the indirect test scored higher on the direct test under social pressure. It only seems counterintuitive, however, under the assumption that all participants have negative implicit attitudes toward Blacks. But a sizable minority of participants showed a pro-Black bias. For these participants, social pressure to avoid race bias meant responding more equally to White and Black faces, resulting in a higher, (i.e., less negative) score on the direct test. For both participants with pro-White biases and those with pro-Black biases, social pressure led to more egalitarian explicit responses, closer to the neutral point of no difference between Black and White faces.

General Discussion

These studies found that the kinds of items people judge and the kinds of judgments they make about them influence whether implicit and explicit evaluations will be related. These factors, among others, make up the structure of a test. Most implicit and explicit tests have very different structures, which may be one reason that they are often weakly correlated. Many of the differences between implicit and explicit tests, however, are unrelated to the conceptual distinctions between implicit and explicit cognition. The result is a confounding of test features with the concepts they are meant to measure (Payne, 2001, 2005; Payne, Jacoby, & Lambert, 2005).

When a direct test with complex questions and Likert scales fails to correlate with an indirect test using reaction times to simple words or pictures, what are we to conclude? The most common conclusion is about distinctions between implicit and explicit cognition. But there are many other possible distinctions based on

these test structures that are overlooked. The present studies suggest that these structural differences have important consequences. We found that tests with poorly matched structures are likely to underestimate the implicit–explicit correlation. But when implicit and explicit tests equate these features, the correlations can be quite high, even on the topic of racial attitudes.

In Studies 1 and 2, we introduced a novel method for measuring indirect and direct ratings while holding all else constant. Using that method, we compared implicitly measured responses to Black and White faces with explicit attitude tests that varied in structural fit. When comparing indirect ratings with traditional racial-attitude questionnaires, the correlations were low. But the correlations were much higher when the tests were equated in structure. In Study 3, we provided evidence that these correlations were not artificially inflated by shared method variance. Structural similarity also affected correlations between different explicit measures, a result that was originally found years ago (Ajzen & Fishbein, 1977). But the novel finding that structural differences can obscure implicit–explicit correlations and thus alter conclusions about implicit cognition has not been widely appreciated.

Structural differences are not the only reason that implicit and explicit tests diverge. In Study 4, we tested the role of social pressure in controlling prejudice, one of the most common reasons for using implicit tests. Introducing social pressure produced more egalitarian responses on direct ratings, diminishing the correlation with indirect ratings. This pattern provides further evidence that direct and indirect ratings made on the same items differ in the degree to which they allow intentional control. Even when structurally matched, direct and indirect tests provided unique information. Because they were structurally matched, it was easier to pinpoint the sources of difference between the two tests.

Test Sensitivity and the Ubiquity of Implicit Bias

The degree of race bias on indirect AMP ratings varied across our samples, sometimes showing a significant bias on average and in other samples showing no significant bias. To examine this issue more closely, we calculated the proportion of participants showing a relative preference for White faces over Black faces on the indirect AMP ratings across all four studies. For White participants ($n = 184$), 54% showed more favorable indirect ratings for White trials than Black trials, whereas 46% showed no difference or a pro-Black preference. For Black participants ($n = 41$), there was a pro-Black bias on average, with 61% showing a pro-Black bias and 39% showing no bias or pro-White bias. The difference for White participants was significant when aggregated across samples, but the difference for Black participants, given the small sample, was not. The relatively small average bias might be seen as evidence that the AMP was not sensitive to White participants' racial attitudes. That interpretation, however, rests on the assumption that our participants' true automatic evaluations were strongly anti-Black. It is also possible that their attitudes were genuinely closer to neutral, on average, and that the AMP accurately detected this.

Previous research using the implicit association test (Greenwald et al., 1998), for instance, has shown that large majorities of Americans tested show anti-Black bias (Nosek, Banaji, & Greenwald, 2002). Some researchers, however, have argued that this large average bias is inflated by other factors such as "extrapersonal" associations (Olson & Fazio, 2004; Karpinski & Hilton, 2001). Research using other methods, such as evaluative priming, has shown that only about half of the individuals tested showed anti-Black bias (Fazio & Olson, 2003). Our results using the AMP suggest a slight majority of White participants in our college samples showed anti-Black bias. None of these figures, however, is very meaningful by itself. Effect size is an indicator of sensitivity only if the effect size is systematically related to the strength of the underlying attitude. A measure that shows large effect sizes on average without the ability to predict individual differences in behavior would not provide evidence of sensitivity.

The correlations demonstrated here provide evidence of sensitivity. In Studies 1 and 2, the positive correlations suggested that individuals who have strong preferences for White faces over Black faces do show large indirect AMP effects. This relationship can be seen most clearly when there is little motivation to avoid expressing prejudice, as shown in Study 4. Together, these data suggest that although anti-Black bias may be small on average (in college samples), the methods presented here capture bias effectively.

Defining Implicit Cognition

An influential way to define implicit social cognition has been offered by Greenwald and Banaji (1995), who defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8). At the core of this conceptual definition are unobservable entities, "traces," that are said to mediate outwardly observable behaviors. But measurement can reach only the observable behaviors, leaving loose connections between conceptual and operational definitions, and plenty of room for controversies over the meaning of implicit tests. The AMP with direct and indirect instructions defines implicit evaluations as those evaluations that influence responses despite the intention to avoid that influence. In contrast, it defines explicit evaluations as those that are intentionally expressed. The advantage is that this method links operational and conceptual definitions in an empirically verifiable way while holding all else constant, in accordance with the "intentionality criterion" of Schacter et al. (1989). Explicit attitudes are those expressed intentionally; implicit attitudes are those expressed despite intentions.

Implications for Theories of Implicit Attitudes

Theories of implicit attitudes have been developed mainly to explain dissociations between implicit and explicit measures. Reviews of the literature suggest that the average correlation between implicit and explicit attitude tests is quite low (Blair, 2001; Fazio & Olson, 2003; Hofmann et al., 2005). In the context of such findings, it stands to reason that theories would focus on the differences between implicit and explicit attitudes. However, the present research demonstrates that the correlation greatly increases when extraneous differences between implicit and explicit measures are eliminated. In light of these findings, future theories might take a different direction.

Consider for example, how theories of implicit attitudes might change if in a future meta-analysis, researchers found, using structurally fitted methods, that the average correlation was actually .70. This may not be a far-fetched scenario, as we found in the

present studies strong correlations when features were matched, and the topic of racial attitudes typically elicits implicit–explicit relationships that are lower than for many other topics. When examined across many topics, such a high average correlation may be realistic. Faced with high correlations, theorists might place a greater emphasis on the similarities between implicit and explicit attitudes. Those instances when implicit and explicit attitude tests diverge would become the exception to be explained by specific circumstances, rather than the rule.

The results of our studies do not mean, however, that implicit and explicit attitudes are identical. It is difficult to explain some research findings with method differences alone. For example, our studies showed that social pressure uncoupled the two tests even when they were equated in structure and that they correlated differently with person judgments. Other research shows that implicitly and explicitly measured attitudes may change at different rates (Gregg, Seibt, & Banaji, 2006; Petty, Tormala, Briñol, & Jarvis, 2006) and that divergence between the two may lead people to seek out more information (Brinol, Petty, & Wheeler, 2006). These results provide evidence that the constructs being tapped by the implicit and explicit tests used in that research can have different determinants and different consequences. But even here, the conclusion that implicit and explicit *cognition* therefore have these different determinants and consequences relies on the assumption that the only difference between tests is the difference between implicit and explicit cognition. If the tests differ in multiple ways, then the differences observed might be due to any of them. Tests with high structural fit could help narrow the options to isolate the mechanisms that are most critical in this important research.

Even when implicit and explicit tests are correlated in the .50 range (as they were in our studies, mean $r = .53$ for structurally equated tests), there remains a great deal of variance for which researchers must account. As the correlation between implicit and explicit tests grows larger, however, the need for theories to explain divergence becomes smaller, and the need to explain similarities becomes greater. It is therefore important to sort out when low correlations between implicit and explicit measures reflect real differences between accurately measured constructs and when they reflect stronger correlations obscured by methodological differences. In many cases, researchers want to ask questions specifically about those situations in which implicit and explicit evaluations differ. In those cases especially, it is important to hold irrelevant aspects of attitude measurement constant so that dissociations can be traced to psychological differences rather than to method differences.

Flaw or Feature?

We have treated many of the differences between implicit and explicit tests as flaws to be fixed to more cleanly make distinctions between implicit and explicit cognition. Readers may find some of these differences to be valuable features instead. For example, researchers might be interested in the different levels of complexity involved in responding to explicit attitude questionnaires versus simple items on an implicit test. Or the slow presentation of questionnaire items versus the fast presentation of priming stimuli might be of theoretical interest. One person's feature is another's flaw, and the answer depends on one's purpose.

We agree that such differences are potentially interesting and important. Our argument is *not* that implicit and explicit tests should always be equated on these features, regardless of the research question. Our argument is that more will be learned if researchers design implicit and explicit tests that differ specifically on those features that are relevant for their particular question, in the same way that experimental manipulations are targeted to answer a specific question.

Mental processes can be automatic or controlled in a variety of senses (Bargh, 1989), and the same applies to implicit versus explicit attitude tests (De Houwer & Moors, 2007). In this article, we have focused on intent, one feature essential to the implicit–explicit distinction. There are, of course, other important distinctions that researchers often wish to draw. For example, a researcher might wish to present the attitude objects slowly in the direct test to encourage participants to think elaborately about them. Of course, differences could be driven by intent, by processing time, or by both. These explanations could be teased apart, if desired, by varying them independently. Our recommendation on the basis of this research is that researchers should identify those cognitive processes that are most important to the research question (such as intent, elaboration, awareness of the attitude object, etc.) and manipulate those features. That requires some theory-guided choices about what factors researchers expect to make a difference and why. Although many measurement techniques do not allow for this possibility, we have proposed a method that allows researchers to manipulate many differences between implicit and explicit tests, one feature at a time. The critical point is that each important feature can be deliberately manipulated to test hypotheses, without confounding these features with extraneous differences. Doing so not only provides greater methodological precision but also encourages researchers to specify what mental processes they refer to when using terms such as "implicit" and "explicit," as has been suggested by many authors (Bargh, 1989; Fazio & Olson, 2003; De Houwer & Moors, 2007).

Conclusion

Sometimes differences between test structures are mistaken for differences between implicit and explicit thought. We suggested one means for equating test structures to solve that problem. When the tests were equated, much of the divergence between them evaporated, leaving implicit and explicit tests highly correlated. But even when equated in structure, direct and indirect tests did not agree perfectly. A snug structural fit still leaves room for the kinds of differences predicted by dual process theories. In fact, those remaining divergences are all the more interesting because method-based explanations can be ruled out with greater confidence. As with microscopes and telescopes, implicit tests require delicate focusing. The more closely the tests are aligned, the easier it is to see their similarities. And that makes it easier to focus clearly on their differences.

References

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888–918.

Arkes, H., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or

would Jesse Jackson fail the implicit association test? *Psychological Inquiry, 15,* 257–278.

Bagozzi, R. P., & Yi, Y. (1991). Multitrait–multimethod matrices in consumer research. *Journal of Consumer Research, 17,* 426–439.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influences in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New York: Guilford Press.

Blair, I. (2001). Implicit stereotypes and prejudice. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition* (pp. 359–374). Mahwah, NJ: Erlbaum.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Brigham, J. C. (1993). College students' racial attitudes. *Journal of Applied Social Psychology, 23,* 1933–1967.

Brinol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. *Journal of Personality and Social Psychology, 91,* 154–170.

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

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

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

De Houwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 179–194). New York: Guilford Press.

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

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,* 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.

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

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

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90,* 1–20.

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

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

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

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin, 27,* 212–225.

McConahay, J. B. (1983). Modern racism and modern discrimination: The effects of race, racial attitudes, and context on simulated hiring decisions. *Personality and Social Psychology Bulletin, 9,* 551–558.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology, 64,* 723–739.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes and Intergroup Relations, 8,* 39–52.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics, 6,* 101–115.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the implicit association test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86,* 653–667.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181–192.

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive control modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology, 89,* 488–503.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293.

Payne, B. K., Govorun, O., & Arbuckle, N. L. (in press). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition and Emotion.*

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2005). Attitudes as accessibility bias: Dissociating automatic and controlled components. In R. Hassin, J. Bargh, & J. Uleman (Eds.), *The new unconscious* (pp. 393–420). Oxford, England: Oxford University Press.

Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST Model. *Journal of Personality and Social Psychology, 90,* 21–41.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043–1056.

Schacter, D. L., Bowers, J., & Booker, J. (1989). Intention, awareness, and implicit memory: The retrieval intentionality criterion. In S. Lewandowsky, J. C. Dunn, & K. Kirsner (Eds.), *Implicit memory: Theoretical issues* (pp. 47–65). Hillsdale, NJ: Erlbaum.

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 336–342.

Warrington, E. K., & Weiskrantz, L. (1974). The effect of prior learning on subsequent retention in amnesic patients. *Neuropsychology, 12,* 419–428.

Wicker, A. W. (1969). Attitudes vs. actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues, 25,* 41–78.

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

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72,* 262–274.

Received May 8, 2006
Revision received July 6, 2007
Accepted July 9, 2007 ■