The Reliability of Implicit Stereotyping

Kerry Kawakami
University of Nijmegen

John F. Dovidio
Colgate University

Recent research has moved beyond the mere documentation of implicit stereotypes to consider how these measures relate to attitudes and predict behaviors. Little is known, however, about the basic psychometric properties of these measures. The present research includes three studies that provide evidence for test-retest reliability of implicit stereotypes when supraliminal priming of associated traits precedes a group categorization decision (Experiments 1 and 2) and when subliminal presentation of a group member precedes a decision about trait applicability (Experiment 3). Across the studies, significant evidence of implicit racial and gender stereotyping was obtained. These effects showed moderate test-retest reliability of comparable levels from 1 hour to 3 weeks. Implications of these findings for the use of implicit measures are considered.

Response latency procedures and other techniques, often borrowed from cognitive psychology, have been frequently used in social psychology to assess the content of stereotypical representations (Banaji & Greenwald, 1995; Dovidio, Evans, & Tyler, 1986; Gaertner & McLaughlin, 1983; Hense, Penner, & Nelson, 1995) and evaluative associations, attitudes, and prejudices (Dovidio & Fazio, 1991; Fazio, Jackson, Dunton, & Williams, 1995; Fazio, Sanbonmatsu, Powell, & Kardes, 1986). These techniques potentially assess implicit activations and offer a conceptually and empirically different perspective on both stereotypes and attitudes than traditional self-report measures.

The distinction between explicit and implicit memory processes has recently received substantial empirical attention (e.g., Loftus & Klinger, 1992; Schacter, 1990). Whereas implicit memory processes involve lack of awareness and automatic activation, explicit processes are conscious, deliberative, and controllable. Demonstrations of dissociations between implicit and explicit cognition, for example, come from studies showing that specific experimental variables produce different and even opposite effects on explicit and implicit tasks (Murphy & Zajonc, 1993; Richardson-Klavehn & Bjork, 1988). A similar distinction has emerged in the literature on stereotyping and attitudes. Specifically, Greenwald and Banaji (1995; Banaji & Greenwald, 1995) have emphasized the importance of distinguishing between explicit and implicit indices of stereotyping and attitudes. Explicit measures of stereotypes and attitudes operate in a conscious mode and are exemplified by traditional measures of these constructs (e.g., Fishbein & Ajzen, 1975; Katz & Braly, 1933). Implicit stereotypes and attitudes, in contrast, operate in an unconscious fashion. Implicit stereotypes are "introspectively unidentified (or inaccurately identified) traces of past experience that mediate attributions of qualities to members of a social category" (Greenwald & Banaji, 1995, p. 15), and implicit attitudes are "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8).

Research using a range of response latency procedures has demonstrated that stereotypes may operate like other semantically related concepts (e.g., doctor-nurse) (Meyer & Schvaneveldt, 1971) to facilitate responses and decision making. Gaertner and McLaughlin (1983), for example, used a lexical decision task in which participants were asked to make judgments about whether pairs of letter strings presented simultaneously were both words (Meyer & Schvaneveldt, 1971). They

Authors' Note: We are grateful for the helpful comments and suggestions offered by Jerry Suls and the anonymous reviewers. The work in this article was supported, in part, by National Institute of Mental Health Grant MH 48721 to the second author. Correspondence should be addressed to Kerry Kawakami, Department of Social Psychology, University of Nijmegen, Postbus 9104, 6500 HE Nijmegen, the Netherlands; email: kawakami@psych.kun.nl.

PSPB, Vol. 27 No. 2, February 2001 212-225

found that both high- and low-prejudiced individuals made their decisions about letter strings faster when the words *Blacks* (or *Negroes*) and *Whites* were paired with stereotype-consistent than with stereotype-inconsistent words. Studies by Dovidio et al. (1986), Baker and Devine (1988), and Zárate and Smith (1990) using different paradigms offer generally convergent results.

Although the above studies demonstrate that people can respond faster to semantically related social stimuli (i.e., stereotypes) than to semantically unrelated social stimuli, they do not necessarily demonstrate preconscious or automatic processes (Bargh, 1994; Greenwald & Banaji, 1995; Kihlstrom, 1990). More recent research specifically designed to elicit automatic responses, however, has demonstrated automatic stereotypic activation toward Blacks (Devine, 1989; Kawakami, Dion, & Dovidio, 1998; Lepore & Brown, 1997; Wittenbrink, Judd, & Park, 1997), women and men (Banaji & Greenwald, 1995; Banaji & Hardin, 1996; Banaji, Hardin, & Rothman, 1993; Blair & Banaji, 1996), elderly people (Hense et al., 1995; Perdue & Gurtman, 1990), Asians (Macrae, Bodenhausen, & Milne, 1995), and a variety of other social categories such as skinheads (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000), soccer hooligans, child abusers (Macrae, Stangor, & Milne, 1994), and professors (Dijksterhuis & van Knippenberg, 1996).

In general, the distinction between implicit and explicit measures of stereotypes and attitudes proposed theoretically by Greenwald and Banaji (1995) is reflected empirically in the literature. Implicit and explicit measures of stereotypes are only weakly and inconsistently related (see Blair, 1999). For instance, Banaji and Hardin (1996, Study 1) reported weak relationships between an implicit measure of gender stereotyping (in which masculine and feminine roles were used to prime decisions about male and female pronouns) and explicit gender stereotypes ($r = .04$). They also found a weak relationship between implicit gender stereotyping and explicit gender attitudes ($r = -.05$). Kawakami et al. (1998) found a correlation of .25 between an implicit category priming measure (in which category labels Black and White were used to prime responses to stereotypic words) and explicit endorsement of racial stereotypes.

Recent research, particularly in the area of attitudes, further indicates that implicit and explicit measures may be tapping different aspects of orientations toward others and their expression (e.g., the role of social desirability in responding) (see Dovidio & Fazio, 1991) and thus may predict different types of behavior. Dovidio, Kawakami, Johnson, Johnson, and Howard (1997), for instance, proposed that implicit (unconscious) aspects of racial attitudes are better predictors of spontaneous behaviors, whereas explicit measures of racial attitudes are better predictors of behaviors in situations in which social desirability factors are salient. Consistent with this reasoning, Fazio et al. (1995) found that direct ratings concerning the legitimacy of the Rodney King verdict and the illegitimacy of the anger of the Black community were correlated mainly with self-reported prejudice (Modern Racism). These responses did not correlate with the response latency measure. However, the response latency measure correlated more highly with the relative responsibility ascribed to Blacks and Whites for the tension and violence that ensued after the verdict than did the Modern Racism scores, the former measure being perhaps a more subtle and indirect manifestation of racial bias. Consistent with these findings, Dovidio et al. (1997) found that self-reported (explicit) racial attitudes primarily predicted overt evaluations of Black and White interaction partners, whereas the response latency measure of implicit attitudes primarily predicted differences in nonverbal behaviors (blinking and visual contact). Thus, response latency measures promise to offer a valuable complement to explicit, self-report measures.

If response latency measures of either attitudes or stereotypes are to be used to predict behaviors or other responses, then it is important that the psychometric properties of these measures be examined. For instance, in their edited book *Measures of Personality and Social Psychological Attitudes*, Robinson, Shaver, and Wrightsman (1991) outline basic evaluative criteria for these measures that include "reliability (both test-retest reliability and internal consistency) and validity (both convergent and discriminant)" (p. 2). Although explicit measures of stereotyping and prejudice often include a detailed report on tests related to the reliability and validity of the scale (for a recent example, see Glick & Fiske, 1996), researchers often fail to justify the use and development of implicit measures related to these topics (Brauer, Wasel, & Niedenthal, 1999). Although an initial attempt has been made to examine the convergent validity of a number of measures of implicit stereotyping (Brauer et al., 1999), to our knowledge no one has yet examined the reliability of response latency measures of attitudes or stereotypes.

The main objective of the present research was to take an initial step toward filling this void. Specifically, we examined the test-retest reliability of two different techniques that have been used for assessing implicit stereotypes. One paradigm uses characteristics as primes and involves making group categorization decisions (Banaji & Hardin, 1996); the other paradigm uses group categories as primes and involves making decisions about stereotypic characteristics (Dovidio et al., 1997). Theoretically, one explanation for the weak and highly variable relationships obtained in the literature between implicit and explicit measures may be limitations in the reliabil-

ity of the implicit measures. Thus, the present research has theoretical as well as methodological implications.

Besides examining the test-retest reliability of implicit stereotyping measures, a secondary objective of the current work was to conceptually replicate earlier investigations using images of actual category members as representations of social categories. Whereas previous research showing implicit stereotyping relies primarily on category labels (e.g., the word *Blacks*) (Wittenbrink et al., 1997) to represent social groups, the present research used photographs (Studies 1 and 2) and schematic faces (Study 3) as stimuli. These stimuli offer a more ecologically valid test of the automatic activation of stereotypes. Theoretically, however, it is possible that pictures and category labels may not produce the same results. Because category labels are notably devoid of any other information except the social category, they ensure that people will respond on the basis of category membership. Images of actual people who belong to the group, alternatively, are more externally valid because similar to all social category members, they may elicit responses to specific, potentially idiosyncratic facial cues (Zebrowitz, Montepare, & Lee, 1993) or produce more individualized rather than category-based responses (Brewer, 1988; Fiske & Neuberg, 1990). It is conceivable that these factors could mitigate the automatic activation of category-based stereotypes. The present work, therefore, potentially contributes theoretically to an understanding of the generalizability or limitations of previous research that used category labels of social categories as primes to elicit implicit stereotypes.

In summary, the present research consisted of three studies of implicit stereotyping that explored the test-retest reliability of response latency measures of implicit gender stereotyping (Experiment 1) and racial stereotyping (Experiments 2 and 3) for two different tasks employed in previous research. One task, based on the work of Banaji and Hardin (1996), used traits as supraliminal primes for categorization decisions about photographs on the basis of sex (Experiment 1) or race (Experiment 2). The other task, based on research by Dovidio et al. (1997), employed subliminal racial category primes (schematic faces) for decisions about traits.

EXPERIMENT 1

The first study examined implicit gender stereotyping and the test-retest reliability of two tasks closely related in time. Specifically, Experiment 1 adapted a supraliminal priming paradigm used by Banaji and Hardin (1996) and Blair and Banaji (1996) in which stereotypic words are used as primes. With this technique, participants are presented with two stimuli sequentially and asked to make a simple decision about the target stimulus (e.g., "Is this a male or female pronoun?"). To create conditions for demonstrating automaticity, the time between the presentation of the prime and the presentation of the second stimulus (stimulus onset asynchrony [SOA]) is short. With short SOAs, participants are unable to engage, focus, and commit attention intentionally and therefore are assumed to be unable to control their response (Neely, 1977, 1991).

As in the Banaji and Hardin (1996) and Blair and Banaji (1996) studies, the primes in the present study were stereotypic words but the target stimuli were photographs of male or female college students. The participant's task was to indicate whether the person in the photograph was a man or a woman. It was hypothesized, as in the Banaji and Hardin (1996) study, that gender stereotypes, even when presented as unrelated to the task, would facilitate categorization of the members of the corresponding sex. Specifically, an overall implicit gender stereotyping effect would be reflected by a Stereotypicality of Prime Word × Sex of Stimulus Face interaction. Female stereotypes were expected to facilitate the categorization of female faces relative to male faces, whereas male stereotypes were predicted to facilitate categorization of male faces relative to female faces. An overall index represented by the contrast for these effects (+1, –1, –1, +1) (see Dovidio et al., 1997; Wittenbrink et al., 1997) was computed for each participant as a measure of individual differences in implicit stereotyping. Conceptually, these coefficients directly represent the interaction of prime word and stimulus face. In other words, the score produced by this linear combination is an index of the extent to which participants responded faster when categorizing female than male faces following female stereotype primes (+1, –1) and responded faster when categorizing male than female faces following male stereotype primes (–1, +1). Participants' explicit attitudes also were assessed using Swim, Aiken, Hall, and Hunter's (1995) Modern Sexism Scale.

To examine test-retest reliability, participants performed the task twice in the same session. The general implicit gender stereotyping effect (i.e., a Stereotypicality of Prime Word × Sex of Stimulus Face interaction) was expected to be consistent across the two administrations of the task. The test-retest reliability was assessed by the Pearson correlation coefficient between the implicit stereotyping scores for each participant for the two categorization tasks.

Method

Participants. Participants were 11 White male and 31 White female undergraduates from a university in the Netherlands. For their involvement in the study,

participants received 10 Dutch guilders (approximately U.S.$6).

Procedure. The study consisted of four phases. The first phase involved a timed person categorization task performed on a PowerMac microcomputer, the second phase consisted of a filler questionnaire that assessed the participants' current emotional state, the third phase was once again the timed person categorization task, and the last phase involved a questionnaire that assessed self-reported sexist attitudes.

Participants were informed by a White female experimenter that the study examined how individuals categorize people. Specifically, the categorization task was described as an "experiment that concerns your speed and accuracy in identifying photographs" under different conditions. Participants were informed that they were in a condition involving distracter stimuli. Furthermore, they were told that before each photograph a distracter word that was unrelated to the categorization task would "appear for a short time" and they were asked to read the word silently. Then, when the photograph appeared on the screen, they were instructed "to judge as quickly and accurately as possible whether the person in the photograph is a man or a woman and to press the appropriate key on a button box." The button box was placed directly in front of the participant, and the meaning of the response keys was counterbalanced across participants.

Specifically, the beginning of each trial was signaled by an asterisk (*) presented in the center of the screen for 500 ms. The asterisk was immediately followed by a trait prime presented for 250 ms. Next, a blank screen appeared for 50 ms before the onset of the photograph. Thus, the interval between the onset of the word prime and the onset of the target photograph was 300 ms. Such short SOAs have been identified as a parameter for eliciting automatic (vs. controlled) responses with supraliminal presentations (Blair & Banaji, 1996; Neely, 1977, 1991). These 4.85 × 3.75-in. photographs, which were obtained with permission from college yearbooks, were presented until the participant responded.[1] In accordance with strategies used by Kawakami et al. (1998) to stimulate participants to respond quickly, participants were presented with their own response latency for 1,000 ms. Last, a blank screen was presented for 750 ms before the next trial.

The trait primes for the categorization task included eight positive and eight negative stereotypes for men and women. Stereotypes were selected on the basis of a pilot study that indicated that these traits were associated more ($ps < .05$) with either men or women. Given those constraints, the traits were further matched on valence and word length. The positive male stereotypes in this study translated from Dutch were *direct*, *brave*, *athletic*, *powerful*, *technical*, *straightforward*, *practical*, and *enterprising*; the negative male stereotypes were *messy*, *macho*, *dominant*, *reckless*, *closed*, *loud*, *aggressive*, and *blunt*. The positive female stereotypes were *caring*, *fashionable*, *clean*, *considerate*, *sociable*, *emotional*, *shy*, and *nurturing*; the negative female stereotypes were *naive*, *unsure*, *gossipy*, *jealous*, *talkative*, *complaining*, *fickle*, and *dependent*. Furthermore, 16 filler traits were used to make the cover story more plausible and to divert the participants' attention from the true purpose of the task by diluting the percentage of gender-stereotypic traits. In accordance with an earlier study by Macrae et al. (1995), these filler words consisted of words unrelated to either male or female stereotypes. Specifically, 8 positive and negative traits that were not differentially associated with men or women ($t < 1.00$) were matched on valence and word length and included in the study. These filler traits were *loyal*, *generous*, *honest*, *happy*, *musical*, *precise*, *optimistic*, *sensible*, *mean*, *greedy*, *strict*, *conceited*, *sneaky*, *cheap*, *unreasonable*, and *conservative*.

Two blocks of trials were presented for each of the categorization tasks. Each block consisted of 48 trials in which each positive stereotype, negative stereotype, and filler trait was presented once. Within a given block, 24 photographs of White men and women were presented. Across blocks, pictures of men and women were both paired once with the 8 positive and 8 negative male and female stereotypes and the 16 filler traits, resulting in a total of 96 trials. Over all trials, each photograph was presented only once; each trait was presented twice, paired once with a male photograph and once with a female photograph. Participants were given rest periods at the end of each block and were instructed to indicate when they were ready to proceed with the study. Before the experimental trials, participants were presented with a practice block of 12 trials. These trials included six photographs of men and women, two male and female stereotypes, and 2 filler traits not used in the actual experiment.

The participants' intermediate task, which separated the test and retest phases of the categorization task, was to estimate their average response latency per trial from the first task and to respond to questions about their current emotional state. The latter questionnaire was presented as a separate activity not related to the timed categorization task and lasted approximately 15 minutes. After completing the second phase, participants performed the categorization task a second time. The random order of the stimulus trials in the second administration of the categorization task was different from the first administration. In the last phase of the study, the participants answered eight questions from Swim et al.'s (1995) Modern Sexism Scale. The internal consistency of these items measured by Cronbach's alpha was .79.

Results and Discussion

The first set of analyses examined whether the implicit priming effect, reflected by a Stereotypicality of Priming Word × Sex of Stimulus Face interaction, would be obtained for gender stereotypes (Banaji & Hardin, 1996; Blair & Banaji, 1996) and whether this effect would be similar across the test-retest administrations of the categorization task. The second set of analyses directly assessed the reliability of responses in the two administrations of the categorization task. The third set explored the relationship between sexist attitudes and stereotypic biases.

Implicit gender stereotyping. Few errors in categorizing faces as male or female were made in the first administration of the task (6.1%) and in the second administration of the task (6.9%). A 2 (stereotypicality of prime word) × 2 (valence of prime word) × 2 (sex of stimulus face) × 2 (timing of task) repeated-measures analyses of variance was conducted on the errors in categorizing faces. Because there were no systematic effects in the present analyses of the errors or the subsequent analyses of the response latencies, sex of participant was not included as an independent variable. Specifically, the three-way Priming Word × Sex of Stimulus Face × Participant Sex interactions for errors and response latencies did not approach statistical significance (Fs < 1).

In accordance with the hypothesized facilitation effects for response latencies for stereotype priming, a significant Stereotypicality of Prime Word × Sex of Stimulus Face interaction was found on the errors in categorizing faces, $F(1, 41) = 18.78$, $p < .001$. Simple effects analyses demonstrated that whereas male stereotype priming produced fewer categorization errors of male faces (M = 2.8%) relative to female faces (M = 9.9%), $F(1, 41)$ = 23.52, $p < .001$, female stereotype priming produced somewhat fewer categorization errors of female faces (M = 5.6%) relative to male faces (M = 9.1%), $F(1, 41)$ = 3.69, $p < .06$.

Before analyzing the response latencies, responses associated with errors and outlier latencies of 3 or more standard deviations beyond each participant's mean (1.3% for the first administration; 1.8% for the second administration) were excluded from the analysis. Remaining response times were subjected to logarithmic transformations (see Blair & Banaji, 1996; Ratcliff, 1993). The primary analysis compared male stereotypes and female stereotypes as primes. For these analyses, the transformed values associated with each of the eight primes were averaged within the four Stereotypicality (male stereotype, female stereotype) × Valence (positive, negative) conditions separately for male and female photographs. All of the analyses were performed on the transformed data, but the untransformed means (in ms) are reported in the text.

Implicit gender stereotyping was assessed by the pattern of responses to the categorization task across the two administrations of the task. A 2 (stereotypicality of prime word) × 2 (valence of prime word) × 2 (sex of stimulus face) × 2 (time of task) repeated-measures analysis of variance was performed on the transformed response latencies. This analysis revealed a main effect for time of task, $F(1, 41) = 10.74$, $p < .01$. Overall, participants' responses were faster during the second administration (M = 403) of the task than the first (M = 415). The predicted Stereotypicality of Prime Word × Sex of Stimulus Face interaction also was obtained, $F(1, 41) = 75.97$, $p < .001$. As expected, male stereotypes significantly facilitated categorization of male faces (M = 395) relative to female faces (M = 423), $F(1, 41) = 27.77$, $p < .001$. Alternatively, female stereotypes facilitated the categorization of female faces (M = 394) relative to male faces (M = 424), $F(1, 41) = 34.69$, $p < .001$.

These effects were consistent across the two administrations of the task; the Stereotypicality of Prime Word × Sex of Stimulus Face × Time of Task interaction did not approach significance, $F(1, 41) = 2.01$, $p = .16$. Similar patterns were obtained across the first (male stereotype/male face = 402, female stereotype/male face = 432, male stereotype/female face = 424, female stereotype/female face = 402) and second administrations of the task (male stereotype/male face = 388, female stereotype/male face = 415, male stereotype/female face = 423, female stereotype/female face = 387).[2] Thus, the results for stereotypicality were consistent with predictions and stable across the two administrations of the task.

Reliability. Test-retest reliability was assessed using the Pearson correlation coefficient for the overall index of stereotypic responding that represented the degree to which stereotypes facilitated categorization of associated members. Specifically, this measure, which was computed separately for the first and second administration of the task for each participant, represented the linear combination of the extent to which male stereotypes produced faster categorization of male relative to female faces and the extent to which female stereotypes produced faster categorization of female relative to male faces (i.e., the single degree of freedom contrast: +1, –1, –1, +1). The reliability of responses, based on the Pearson correlation coefficient (40 df) was .56, $p < .001$. Thus, test-retest reliability, at least for administrations close in time, showed moderate reliability.

Explicit attitudes. Although not a focus of the present research, in supplementary analyses we explored the relationship between explicit attitudes and implicit

stereotyping as assessed by the correlation between scores on Swim et al.'s (1995) Modern Sexism Scale and the composite measure of implicit stereotyping based on the extent to which female stereotypic words facilitated categorization of female faces relative to male faces and male stereotypic words facilitated the categorization of male faces relative to female faces. Sexism scores were not significantly correlated with implicit stereotyping responses on the first categorization task ($r = .16$), the second categorization task ($r = .12$), or the average of the two categorization tasks ($r = .16$).

EXPERIMENT 2

Whereas Experiment 1 examined the test-retest reliability of two administrations within the same session of the supraliminal priming paradigm used by Banaji and Hardin (1996) and Blair and Banaji (1996) for implicit gender stereotypes, Experiment 2 explored the reliability of implicit racial stereotyping across a longer period of time. Specifically, participants in Experiment 2 performed the priming task for racial stereotypes either twice in the same session, as in Experiment 1, or twice across a 5- to 15-day period. Thus, this study extends Experiment 1 by considering implicit stereotyping for race rather than gender and investigates the stability of the measure across a broader time frame.

Method

Participants. Participants were 17 male and 21 female White undergraduates from a liberal arts college in the Northeastern United States. Involvement in the study partially satisfied one option for students' course requirement. These participants were randomly selected from a pool of 147 students, the majority of whom were administered, along with several other questionnaires, Brigham's (1993) 20-item Attitudes Toward Blacks Scale and a 5-item version of McConahay's (1986) Modern Racism Scale at the beginning of the semester. Of the 38 participants in the present study, 25 completed both the Attitudes Toward Blacks Scale and the Modern Racism Scale. The internal consistency (Cronbach's alpha) for these participants was .86 for the Attitudes Toward Blacks Scale and .74 for the Modern Racism Scale. The two scales were significantly correlated, $r(23) = .57$, $p < .003$.

Procedure. The procedure and task were modeled after those in Experiment 1, except that the study examined implicit racial stereotypes rather than gender stereotypes. As in Experiment 1, this study involved the administration of two timed person-categorization tasks performed on a PowerMac microcomputer. Also as in Experiment 1, some of the participants ($N = 14$) performed the task twice during the same hour session, once at the beginning of the session and once at the end, after completing a series of questionnaires about campus life (for an unrelated study). Twenty-four of the participants, however, completed the two timed, person-categorization tasks in two different sessions, ostensibly for two different studies, that were 5 to 15 days apart.

In the categorization task, the trait primes included 8 positive and 8 negative stereotypes for Blacks, 8 positive and 8 negative stereotypes for Whites, and 16 filler traits. On the basis of results from a pilot study, the stimuli once again were matched primarily on stereotypic association, then, given those constraints, on valence, and finally on word length. Traits that significantly distinguished between Blacks and Whites were selected as stereotypic for each group. Traits that did not differentiate between the two groups were used as filler traits. The positive Black stereotypes were *musical*, *athletic*, *strong*, *colorful*, *muscular*, *humorous*, *religious*, and *rhythmic*. The negative Black stereotypes were *poor*, *loud*, *angry*, *tough*, *bitter*, *hostile*, *unemployed*, and *intimidating*. The positive White stereotypes were *educated*, *patriotic*, *hopeful*, *wealthy*, *ambitious*, *practical*, *trusting*, and *industrious*. The negative White stereotypes were *weak*, *boring*, *greedy*, *uptight*, *arrogant*, *gullible*, *conventional*, and *materialistic*. The positive and negative traits used as filler traits were *kind*, *loyal*, *sincere*, *outgoing*, *pleasant*, *friendly*, *independent*, *enthusiastic*, *sad*, *nasty*, *weird*, *lonely*, *confused*, *cautious*, *careless*, and *irresponsible*.

The stimulus presentation in Experiment 2 was the same as in Experiment 1. In particular, in each presentation of the task, participants were given two blocks of 48 trials in which photographs of White and Black men taken from college yearbooks were paired once with the 8 positive and 8 negative Black and White stereotypes and 16 filler traits.

Results and Discussion

Implicit racial stereotyping. Errors in categorizing Black or White faces across the 96 trials in each administration of the task were relatively rare (1.9% of the responses in the first categorization task and 2.6% in the second). A 2 (same vs. different session) × 2 (stereotypicality of prime word) × 2 (valence of prime word) × 2 (race of stimulus face) × 2 (time of task) analyses of variance, with repeated measures on the last four factors, was conducted on the number of errors. In contrast to the results for the first study, perhaps because of the much lower error rate for categorizing faces by race in this study (2.3% overall) than by sex in Experiment 1 (6.5%), the Stereotypicality of Prime Word × Race of Stimulus Face interaction was not significant ($p > .50$).

Before analyzing the response latencies, responses associated with errors and outlier latencies of 3 or more standard deviations beyond each participant's mean

(1.2% for the first administration, 1.9% for the second administration) were excluded from the analysis. Remaining response times were subjected to logarithmic transformations (see Blair & Banaji, 1996; Ratcliff, 1993). Analogous to Experiment 1, the primary analysis compared Black stereotypes and White stereotypes as primes. The transformed values associated with each of the eight word primes were averaged within the four stereotypicality (Black stereotypes, White stereotypes) × valence (positive, negative) conditions separately for Black and White photographs.

Implicit racial stereotyping was assessed by the pattern of responses to the categorization task across the two administrations of the task. Thus, a 2 (same vs. different session) × 2 (stereotypicality of prime word) × 2 (valence of prime word) × 2 (race of stimulus face) × 2 (time of task) mixed analysis of variance, with repeated measures on the last four factors, was performed on the transformed response latencies. This analysis revealed a marginally significant main effect for time of task, $F(1, 36) = 3.30$, $p < .08$. Overall, participants' responses were faster during the second administration of the task ($M = 503$) than during the first ($M = 518$). The predicted Stereotypicality of Prime Word × Race of Stimulus Face interaction also was obtained, $F(1, 36) = 10.22$, $p < .003$. As expected, Black stereotypes significantly facilitated categorization of Black faces ($M = 505$) relative to White faces ($M = 521$), $F(1, 37) = 12.89$, $p < .001$. Although White stereotypes tended to facilitate the categorization of White faces ($M = 505$) relative to Black faces ($M = 511$), this effect was not significant. These effects were consistent across the two administrations of the task and were comparable whether the testing occurred in the same session or several days apart. Similar patterns were obtained across the first (Black stereotype/Black face = 512, Black stereotype/White face = 528, White stereotype/Black face = 519, White stereotype/White face = 512) and second administrations of the task (Black stereotype/Black face = 498, Black stereotype/White face = 514, White stereotype/Black face = 503, White stereotype/White face = 499). Neither the Stereotypicality of Prime Word × Race of Stimulus Face × Time of Task interaction nor the Stereotypicality of Prime Word × Race of Stimulus Face × Session interaction approached significance, Fs < 1, ps > .88.[3] Thus, the results for stereotypicality were consistent with predictions and stable across the two administrations of the task regardless of whether the administrations occurred within the same session or several days apart.

Reliability. Test-retest reliability was assessed using the Pearson correlation coefficient for the overall index of stereotypic responding that represented the linear combination of the extent to which Black stereotypic words produced faster categorization of Black faces relative to White faces and the extent to which White stereotypic words produced faster categorization of White faces relative to Black faces. The reliability of responses, based on the Pearson correlation coefficient (36 *df*), across all participants was .51, $p < .001$. For only participants who completed the task twice in the same session, the correlation coefficient was .42, $p < .001$. For just participants who completed the task twice across a several-day period, the correlation coefficient was somewhat higher (.60, $p < .001$). Thus, test-retest reliability is once again moderate and is highly similar to the results for implicit gender stereotyping in Experiment 1.

Explicit attitudes. The relationship between explicit attitudes and implicit stereotyping again was assessed by the correlation between scores on self-report measures of prejudice, in this case the Attitudes Toward Blacks and the Modern Racism Scales, and the composite measure of implicit stereotyping based on the extent to which Black stereotypes facilitated categorization of Black faces relative to White faces and White stereotypes facilitated the categorization of White faces relative to Black faces. As in Experiment 1, the relationships between racial prejudice and responses on the first categorization task, the second categorization task, and the average across the two administrations of the task were examined for the subset of participants ($N = 25$) who completed the self-report scales 6 weeks prior to participating in the experiment. Attitudes Toward Blacks and Modern Racism were not significantly associated with implicit stereotyping responses on the first categorization task, $r(23) = .29$, $p < .16$, $r(23) = .09$, $p < .66$; the second categorization task, $r(23) = .20$, $p < .35$, $r(23) = .12$, $p < .56$; or the average of the two categorization tasks, $r(23) = .28$, $p < .17$, $r(23) = .12$, $p < .56$. Overall, explicit racial attitudes were found to be only weakly related to implicit racial stereotyping.

EXPERIMENT 3

Experiments 1 and 2 were designed to investigate the test-retest reliability of a response latency measure of implicit stereotyping using the Banaji and Hardin (1996) paradigm. In this paradigm, stereotypic traits were used as priming stimuli and participants were instructed to categorize photographs according to category membership. Experiment 3 examined the test-retest reliability of racial stereotyping over a 3-week period using a modified version of the subliminal priming procedure introduced by Perdue, Dovidio, Gurtman, and Tyler (1990, Experiment 3; see also Dovidio et al., 1997). In the Perdue et al. (1990) experiment, ingroup and outgroup pronoun primes ("we" and "they") were presented very rapidly on a computer screen and then visually masked to prevent participants' awareness of the presence of the prime. The mask was a string of letters designed to cue the cate-

gory "persons" or, in the control condition, "houses." The participants' task was to decide whether the target word that followed could ever describe the cued category, persons or houses. Perdue et al. found that ingroup primes presented outside of awareness facilitated responses, relative to outgroup primes, to positive target words.

The primes in the present experiment were schematic faces of Black and White men and women, which were masked by figures representing the cued categories of persons and houses (see Dovidio et al., 1997; see also Chen & Bargh, 1997). The target words were positive and negative Black stereotypes and White stereotypes. Using subliminal priming techniques and masked racial primes of this type ensures that participants are responding without awareness, suggesting automaticity of activation. Participants in Experiment 3 also completed two explicit (i.e., self-report) racial attitudes measures: Brigham's (1993) Attitudes Toward Blacks Scale and McConahay's (1986) Modern Racism Scale.

It was predicted that, in general, the masked racial primes would facilitate responses of the White participants to stereotype consistent words, producing a Race of Prime × Stereotypicality of Target Word interaction across the two administrations of the task. Specifically, participants were expected to respond faster to Black stereotypes following a Black prime than a White prime and faster to White stereotypes following a White prime than a Black prime. Again, an individual difference measure of implicit stereotyping was created from these analytic contrasts. Test-retest reliability was assessed by the Pearson correlation coefficient between the implicit stereotyping scores for each participant for the two categorization tasks.

Method

Participants. Participants were 5 male and 21 female White undergraduates from a liberal arts college in the Northeastern United States. Involvement in the study partially satisfied one option for students' course requirement.

Procedure. Participants were informed by a White female experimenter that the study examined how individuals categorize people and objects. Target stimuli were a subset of stereotypic traits used by Dovidio et al. (1986) and were supported by pilot studies that indicated that these traits were differentially associated ($ps < .05$) with Blacks and Whites. The stimuli that were selected were matched primarily on stereotypic association, then, given those constraints, on valence, and finally on word length. The positive and negative Black stereotypes in this study were *musical*, *athletic*, *lazy*, and *imitative*. The positive and negative White stereotypes were *ambitious*, *practical*, *conventional*, and *stubborn*.

The main experiment used a procedure that was a variation of a subliminal priming procedure used by Perdue et al. (1990, Experiment 3), in which participants were asked to "quickly and accurately categorize objects and persons." In the present study, participants were told that either the letter string that represented the person category (i.e., P) or the letter string that represented the house category (i.e., H) would be presented on a computer screen and followed by an adjective (i.e., the target words). The responses to the person category were of primary theoretical interest; the house category was used as a control condition so that participants would not always respond affirmatively to the target words. In addition to the stereotype traits, eight words that describe houses but that do not normally describe people (*drafty*, *furnished*, *leaky*, *roomy*, *thatch*, *wooden*, *brick*, and *unfurnished*) also were used as target stimuli.

Preceding the person category (P) and the house category (H), however, participants were presented subliminally with schematic faces of Black and White men (Bargh & Pietromonaco, 1982; Devine, 1989; Perdue et al., 1990). Specifically, two Black and White male faces and two Black and White female faces were systematically constructed with Mac-a-Mug software to be comparable (at least based on self-reported ratings involving 30 White students) in perceived attractiveness, intelligence, friendliness, and likability. These schematic faces were employed and are illustrated in previous research by Dovidio et al. (1997).

For the critical trials, the facial primes were presented parafoveally at a location on the screen such that the center of the word was 3.6 cm to the left or right of the fixation point. Based on pretesting and limited by the hardware used to administer the stimuli (PowerMac 7200—75 MHz), the refresh rate of the monitor resulted in a minimum presentation time of 15 ms and a maximum of 30 ms. This prime presentation time is similar to subliminal priming of photographs of African American and Caucasian faces by Chen and Bargh (1997) using a Gateway 486 computer with a VGA color monitor (13 to 26 ms).

The 2 × 1.75-in. facial primes in the present study were immediately followed in the same location by geometrical figures, a "P" within an oval signifying a person or an "H" within a rectangle representing a house. These geometrical figures were used as visual masks to fully cover the area of the screen occupied by the facial primes. The cued category, which visually masked the facial prime, appeared on the screen for 250 ms, after which the target word (a positive or negative Black or White stereotype or an adjective for a house) was then presented in the same location on the screen. The participants' task was to indicate, by pressing the appropriate key, whether the target word could ever describe a member of the cued category—a person or a house. The locations of the "yes"

and "no" keys (Z and M on the keyboard) were counterbalanced across participants. The target word remained on the screen for 750 ms or until the participant pressed the decision key, after which a blank screen appeared for 1500 ms before the next trial. In accordance with Study 1, exposure times were selected to elicit automatic processes. Whereas SOAs less than 300 ms between the initial facial prime and the target word were used to create conditions requiring efficient processing, subliminal priming was used to establish the automatic criterion of unawareness (Bargh, 1994).[4]

In summary, participants were presented with (a) a subliminal prime (i.e., a Black schematic face or a White schematic face); (b) a cued category (P for a person or H for a house); and (c) a target word that was stereotypic of Whites (e.g., *conventional*), stereotypic of Blacks (e.g., *musical*), or did not commonly describe a person (e.g., *drafty*) and were instructed to respond to whether the target word could ever describe a member of the cued category (i.e., a person or a house). Overall, the experiment consisted of 128 trials; 64 trials were of theoretical interest. In these trials, each of the eight-person descriptive words was paired with one White female, one Black female, one White male, and one Black male face presented once to the left of the fixation point and once to the right of the fixation point. Before beginning the actual experiment, participants were first presented with six practice trials to allow them to become familiar with the task. For each participant's performance on the task, an implicit stereotyping score was computed as the linear contrast (+1, –1, –1, +1) representing the degree to which participants responded faster to Black stereotypic words following the Black prime than the White prime and the degree to which participants responded faster to White stereotypic words following the White prime than the Black prime.

After performing the priming task in the first session, participants were administered an "opinion survey" that assessed racial prejudice using a 14-item version of Brigham's (1993) Attitudes Toward Blacks Scale (Cronbach's alpha = .83) and 6-item version of McConahay's (1986) Modern Racism Scale (Cronbach's alpha = .73). Scores on the Attitudes Toward Blacks and the Modern Racism Scales were highly correlated, $r(24) = .72$, $p < .001$. The same participants were contacted 2 weeks later and informed that they had been randomly selected to participate in another decision-making session. These participants performed the identical priming task used in the first session in a range of 18 to 22 days from their first session.

Results and Discussion

As in the first two experiments, the first set of analyses examined whether the implicit priming effect, in this case reflected by a Race of Prime × Stereotypicality of Target Word interaction, would be obtained and whether this effect would be stable across the test-retest administrations of the priming task. The second set of analyses tested the reliability of responses in the two administrations of the priming task. The third set investigated the relationship between racial attitudes and implicit stereotyping.

Priming task. Response latencies related to errors (i.e., when participants responded "no" to the stereotype target words on a trial in which the person category was cued) for the 64 trials of theoretical interest were low (3.1% in the first task, 1.2% in the second task). A 2 (race of prime) × 2 (stereotypicality of target word) × 2 (valence of target word) × 2 (time of task) repeated-measures analyses of variance was performed on the number of errors. No significant interactions involving the sex of facial primes in the analyses of the errors or in the subsequent analyses of the response latencies were found in Study 3.

In the analysis on the number of errors, only the main effect for time of task was significant, $F(1, 25) = 9.06$, $p < .006$. Participants had a significantly lower error rate the second time they performed the task compared to the first. The Race of Prime × Stereotypicality of Target Word interaction did not approach significance ($p > .75$).

Before analyzing the response latencies, responses associated with errors and outlier latencies for the trials of theoretical interest that were 3 or more standard deviations beyond each participant's mean response latencies (2.2% in the first task, 2.4% in the second) were excluded from the analysis. The remaining response times were subjected to a logarithmic transformation. The transformed values associated with each of the two stimulus words were averaged within the four Stereotypicality × Valence conditions for subsequent analyses.

The log-transformed response latencies were subjected to a 2 (race of prime) × 2 (stereotypicality of target word) × 2 (valence of target word) × 2 (time of task) repeated-measures analysis of variance. In general, participants responded faster to White stereotypes ($M = 732$) than to Black stereotypes ($M = 846$), $F(1, 25) = 54.27$, $p < .001$. Participants also responded somewhat faster on the second administration of the task ($M = 743$) than on the first ($M = 836$), $F(1, 25) = 3.82$, $p < .07$. The predicted Race of Prime × Stereotypicality of Target Word interaction was obtained, $F(1, 25) = 8.73$, $p < .007$. Planned comparisons revealed, as expected, that response times to Black stereotypes were faster following Black primes ($M = 801$) than White primes ($M = 891$), $F(1, 25) = 6.15$, $p < .02$. Although responses to White stereotypes were somewhat faster following White primes ($M = 721$) than Black primes ($M = 743$), this difference was not significant. The Race of Prime × Stereotypicality

of Target Word × Time of Task interaction did not approach significance, $F < 1$. Similar patterns were obtained across the first (Black stereotype/Black face = 856, Black stereotype/White face = 960, White stereotype/Black face = 775, White stereotype/White face = 751) and second administrations of the task (Black stereotype/Black face = 746, Black stereotype/White face = 824, White stereotype/Black face = 710, White stereotype/White face = 692).[5]

Reliability. Test-retest reliability was again assessed using the Pearson correlation coefficient for the overall index of stereotypic responding that represented the degree to which participants responded faster to Black stereotypic words following the Black prime than the White prime and the degree to which participants responded faster to White stereotypic words following the White prime than the Black prime (i.e., the +1, –1, –1, +1 linear combination). This measure was computed separately for the first and second administrations of the priming task for each participant. The reliability of responses, based on the correlation coefficient (24 *df*), was .50, $p < .001$. Thus, test-retest reliability across a 3-week period showed moderate reliability and was comparable to reliabilities demonstrated in Experiments 1 and 2.

Explicit racial attitudes. The relationship between the two explicit measures of prejudice and the response latency measure of stereotyping was tested by examining the correlations between the Attitudes Toward Blacks and Modern Racism scores and the composite measure of implicit racial stereotyping. The correlations of Attitudes Towards Blacks and Modern Racism scores with implicit racial stereotypes were nonsignificant across the responses on the first priming task (*r*s = .20 and .09, respectively), the second priming task (*r*s = –.14 and –.08, respectively), and for the average of the responses on the two types of priming tasks (*r*s = .05 and .01, respectively).[6]

GENERAL DISCUSSION

Recently, research has moved beyond the mere documentation of implicit stereotypes and attitudes to consider how these measures may predict behaviors and opinions, often in ways independent of explicit self-reported attitudes. For example, it has been hypothesized that implicit measures may be better predictors of subtle or spontaneous expressions of bias, whereas explicit measures may be better predictors of blatant and deliberative types of bias (Dovidio & Fazio, 1991; Dovidio et al., 1997; Fazio et al., 1995). Despite the initiatives in this promising direction, little is known about the stability and basic psychometric properties of these implicit measures (Brauer et al., 1999).

The primary contribution of the present research involves new evidence related to the test-retest reliability of measures of implicit stereotyping. Specifically, Experiment 1 revealed that an overall measure of implicit gender stereotyping, assessed using the Banaji and Hardin (1996) technique, had modest reliability (.56) across two tests within the same experimental session. Experiment 2 demonstrated a similar level of reliability (.51) for implicit racial stereotyping with the same task. These findings, however, were obtained across a longer period of time (5 to 15 days). Experiment 3 showed that another index of implicit racial stereotyping based on a variation of the Perdue et al. (1990; see also Dovidio et al., 1997) priming procedure was also moderately reliable (.50) over a 3-week period.

Although these reliabilities are reasonable, they are not especially high relative to many well-established explicit measures. For example, in a follow-up to the present research, using an independent sample of participants ($N = 26$), we assessed respondents' estimates of the percentage of Blacks who possess characteristics identified as being part of the Black stereotype (embedded among nonstereotypical characteristics) at two separate times. The second testing occurred 2 to 3 weeks after the first. Although there was an overall decrease in the percentages reported between the two assessments, *M*s = 59.1% vs. 54.1%, $F(1, 25) = 6.28$, $p < .019$, the test-retest reliability was high, $r(24) = .84$, $p < .001$. In addition, with respect to racial attitudes, we have found that the test-retest reliability of Brigham's (1993) Attitudes Toward Blacks Scale over a 3-week period was .89. However, explicit measures are not necessarily more reliable than implicit measures. The test-retest reliability we obtained for the Modern Racism Scale over 3 weeks was .44, which is comparable to the test-retest reliability of our implicit measures of stereotyping.

One explanation for the moderate levels of reliability we obtained for implicit stereotypes may be related to the limitations of our measurement techniques. Although we obtained similar levels using two separate, very different, techniques, assessment of implicit stereotypes and prejudice in general may be less well-honed than those that have been developed for explicit stereotypes and attitudes. In comparison to the latter type of research, which has been the focus of investigation for more than 70 years, interest in implicit stereotyping is fairly new. Research in cognitive science (Salmon & Butters, 1995; Squire & Kandel, 1999), however, suggests one way of honing this type of research. Specifically, recent findings indicate that there may be various forms of implicit processes that can operate independently from one another. To the extent that an implicit measurement technique taps more than one of these systems, responses may reflect a combination of processes that is less stable

than are responses based on a single system, such as conscious processing is assumed to be. Thus, the development of more reliable techniques for measuring implicit stereotypes and attitudes would profit from a greater understanding of the operation and manifestations of different systems of implicit memory. It also is possible that the moderate levels of reliability obtained in the present studies may be related to the phenomenon itself. Perhaps because implicit responses are more global and less finely differentiated than more deliberative, explicit responses (Kihlstrom, 1990), they may be inherently less stable and thus normally have lower levels of reliability.

So what are the implications of our findings for the usefulness and future of implicit stereotyping? If response latency measures of automatic stereotype activation are to be used in the future to predict stereotype application, outgroup attributions, attitudes toward outgroups, and intergroup behavior, it is clear that knowledge concerning the reliability of these measures is essential. Our results indicate that these measures are reliable, but not at the level that is typically desired and reported for explicit measures. Nevertheless, we caution that they should not be dismissed prematurely. Implicit measures may still be theoretically and empirically useful at this time. For instance, some clinically based projective techniques such as the Thematic Apperception Test (TAT) have similar psychometric properties (e.g., test-retest correlations of .48 to .56) (Lundy, 1985) but have provided valuable insights into human motivation, such as achievement motives. McClelland, Koestner, and Weinberger (1989) distinguished between self-reported and implicit motives (as assessed by the TAT), which do not necessarily correlate with one another. They concluded that self-reported motives primarily predict behavior when incentives, rewards, or social expectations are salient, whereas implicit motives (such as those reflected in projective tests) primarily predict behavior in the absence of external demand. The implicit-explicit distinction may thus be a fundamental one for understanding the dynamics of cognitions, feelings, and motivations.

Besides examining the test-retest reliability of implicit measures, a secondary aim of the present studies was to conceptually replicate previous work on automatic racial and gender stereotypes using two separate paradigms. The results demonstrate that implicit stereotypes were automatically activated by the supraliminal priming of an ostensibly unrelated trait that preceded a group categorization decision (Experiments 1 and 2) (see Banaji & Hardin, 1996) or by a subliminal presentation of a group member that preceded a decision about whether a trait can describe a person (Experiment 3) (see Dovidio et al., 1997). Although not a direct indication of convergent validity because the same participants did not participate in both procedures, Experiments 2 and 3, in which participants from the same population of students were involved in one of the two techniques for assessing implicit racial stereotypes, produced generally comparable results.

On one hand, given that the Banaji and Hardin (1996) and Perdue et al. (1990) techniques have been used before as alternative methods of assessing implicit stereotypes and attitudes, the generally converging results should not be surprising. Variations in procedure have mattered little in prior work. On the other hand, theoretically, this convergence may be seen as quite surprising. Lepore and Brown (1997), for example, distinguished between trait priming (related to the procedure in Experiments 1 and 2) and category priming (related to the procedure in Experiment 3). With respect to category priming, categorizing a person as "Black" may indeed activate related stereotypic traits. However, with respect to trait priming, the presentation of a trait such as "musical" may activate concepts other than racial associations, such as opera singers. Attribute-category links may be weaker than category-attribute links simply because attributes do not necessarily determine category membership (Anderson & Klatzky, 1987; Fiske, Neuberg, Beattie, & Milberg, 1987). Thus, it is theoretically possible, and in fact quite probable, that category and stereotype priming are not entirely equivalent, structurally or functionally. Although a more systematic examination of the relationship between the activation of categories and traits is clearly indicated, our results suggest that not only can social category priming automatically activate associated stereotypes but stereotypic trait priming can activate associated categories.

In contrast to earlier investigations that have largely used category labels to induce implicit stereotyping, the present research also demonstrates that exposure to actual category members also can automatically activate stereotypic associations. These findings demonstrate the generalizability of the elicitation of implicit stereotypes by category labels and offers evidence of the ecological validity of these effects. Furthermore, the present research provides evidence of the automatic categorization of people as men or women or as Black or White (see Fiske, 1998). Although individualized responses to facial cues are possible (see Zebrowitz et al., 1993), our findings suggest that an individual's gender or race also can automatically elicit category-based associations. Thus, the present research contributes to the growing body of evidence identifying the range of stimuli that can elicit implicit stereotypes as well as demonstrates alternative techniques for assessing them.

Although there is generally convergent evidence of implicit stereotyping across our three experiments, we note a general asymmetry of results for implicit Black

and White stereotyping. In Experiment 3, responses to Black stereotypes were significantly faster following Black primes than White primes. Responses to White stereotypes were somewhat but not significantly faster following White primes than Black primes. Perhaps this asymmetry could be attributable to the nature of the particular prime used in these studies of schematic faces. Nevertheless, the same pattern appears when stereotypic traits were used as primes in Experiments 2. Black stereotype primes significantly facilitated categorization of Black photographs relative to White photographs, whereas the effects for White stereotype primes facilitating categorization of White photographs relative to Black photographs did not attain statistical significance. Meta-analytic tests across these two studies of implicit racial stereotypes revealed an overall significant effect for responses indicative of Black stereotype activation (i.e., categorization latencies for Black and White photographs following presentation of Black stereotypes in Experiment 2 and responses to Black stereotypes as a function of schematic face primes in Experiment 3), $z = 3.98$, $p < .001$, M Fisher's $z = .52$, $M\,r = .48$. The meta-analytic test of White stereotype activation only approached significance across the two studies, $z = 1.56$, $p < .12$, and the magnitude of the effect was considerably weaker, M Fisher's $z = .20$, $M\,r = .20$.

Theoretically, this asymmetry may be a consequence of the distinctiveness of the minority group, particularly as perceived by majority group members. Zárate and Sandoval (1995; see also Stroessner, 1996) have proposed that there is a "White male default" cultural value by which others are judged. Deviations from this standard are thus distinctive (see also Levin, 1996). Information about members of salient minority groups is, in turn, processed in terms of prototypical ways that facilitate stereotypical organization and activation, whereas information about majority group members involves mainly exemplar-based processing (Mullen, 1991; Mullen, Rozell, & Johnson, 1996). Supportive of this notion, there has traditionally been greater consensus for Black stereotypes than for White stereotypes (Dovidio, Brigham, Johnson, & Gaertner, 1996; Karlins, Coffman, & Walters, 1969). Although this explanation is generally consistent with the results of implicit racial stereotyping (Experiments 2 and 3), it is not consistent with our results for gender stereotyping (Experiment 1), in which trait priming effects were symmetric for categorizing female and male photographs. One important factor affecting group distinctiveness, however, may be relative group size (Mullen, 1991). Because women and men are equally represented in the society and in fact women were overrepresented in the experimental samples, this category may not be perceived to be distinctive. Nevertheless, because category salience and social context have been demonstrated to influence explicit stereotyping (see Spears, Oakes, Ellemers, & Haslam, 1997), future research might focus productively on how group distinctiveness moderates the activation of implicit stereotypes.

Finally, the present studies also add to the growing body of research showing generally weak relationships between explicit prejudice and implicit stereotyping (for reviews, see Blair, 1999; Dovidio et al., 1996). Specifically, across two studies (Experiments 2 and 3), the implicit racial stereotyping measure was correlated .28 and .05 with the Modern Racism Scale and .12 and .01 with the Attitudes Towards Blacks Scale. In Experiment 1, implicit gender stereotyping was correlated .20 with the Modern Sexism Scale. These correlations are similar to findings by other studies on the relationship between explicit racial attitudes (i.e., Modern Racism scores) and implicit stereotyping (e.g., von Hippel, Sekaquaptewa, & Vargas, 1997, $r = .01$; Wittenbrink et al., 1997, $r = .13$) and between explicit sexist attitudes and implicit stereotyping (e.g., Banaji & Greenwald, 1995, Study 3, $r = .10$; see Blair, 1999, for a review).

A number of factors may contribute to the generally weak relationship between explicit attitudes and implicit stereotyping. First, theoretically, Devine (1989) proposed that whereas Whites generally develop implicit racial stereotypes to a similar degree, under more controlled processing conditions, low-prejudiced Whites attempt to suppress these thoughts, whereas high-prejudiced Whites do not. This dissociation for low-prejudiced Whites limits the relationship between explicit prejudice and implicit stereotyping. Second, because the overall relation between explicit measures of racial stereotyping and prejudice has been empirically shown to be modest with a meta-analytic r of .25 (Dovidio et al., 1996), it is possible that the sample sizes in the present experiments were too small to detect these effects. Third, a number of recent theorists have emphasized the importance of the nature of the implicit and explicit concepts that are being measured (Brauer et al., 1999; Dovidio et al., 1996). For instance, Wittenbrink et al. (1997) found that a measure of implicit evaluative bias correlated .41 with Modern Racism and averaged .28 with four other measures of explicit prejudice. The mean correlation between implicit stereotypic associations and these latter four measures, however, was substantially lower, –.02. Finally, the level of specificity of implicit and explicit measures also may be important to their relationship. In particular, whereas prejudice often is assessed toward a particular group (i.e., Blacks, women), stereotyping is assessed by the differences in activation between two groups (i.e., Blacks and Whites or women and men). In summary, although the present findings suggest a consistent, weak relationship between explicit and implicit measures of prejudice and stereotyping, it is clear that future

research is needed to address the many issues related to this topic.

In conclusion, the present research provides consistent evidence indicating test-retest reliability for implicit racial and gender stereotyping. Across three studies and for two distinct response latency paradigms, moderate reliability was found. Although we acknowledge the relative weakness in this psychometric property, we remain positive with regard to these new techniques. Implicit measures not only provide important information about cognitive processes and initial, potentially automatic, first reactions to category members but also are particularly well-suited to measuring socially sensitive issues, such as those related to prejudice and stereotyping. Because even measures designed to be nonreactive, such as the Modern Racism Scale (McConahay, 1986), seem to be vulnerable to self-presentation and social desirability pressures (Fazio et al., 1995), implicit measures that preclude controlled processing may be uniquely valuable for examining sensitive topics related to intergroup bias (Dovidio & Fazio, 1991). To be able to have faith in the findings related to these new "bogus-pipeline" techniques (Fazio et al., 1995), however, further research related to the reliability and validity of the scales is clearly necessary. The present work represents one step in that direction.

NOTES

1. Stimulus materials for all experiments presented are available from the authors.

2. The correlation between each participant's mean latency score overall and their degree of stereotypic biases, calculated as the linear combination (+1, –1, –1, +1) of the extent to which female stereotypic words facilitated categorization of female faces relative to male faces and the extent to which male stereotypic words facilitated categorization of male relative to female faces, was .25 (40 *df*), $p = .11$.

3. In general, participants who took longer to respond overall exhibited greater stereotypic biases, $r(36) = .32$, $p < .05$.

4. To further ensure that participants were indeed unaware of the facial primes, a pilot guessing study was conducted in which 12 participants were run through a 48-trial procedure similar to the main experiment but were asked to guess the initial prime. Comparable to the rates reported by Bargh and Pietromonaco (1982), Devine (1989), and Perdue, Dovidio, Gurtman, and Tyler (1990) with words as primes, on only 17 of the 576 trials (3%) did these participants accurately identify the prime. These data support the results of the debriefing and indicate that the effects of the facial priming can be considered automatic because processing occurred without conscious awareness.

5. Participants who took longer to respond overall showed somewhat, but not significantly, greater stereotypic biases, $r(24) = .16$, $p < .43$.

6. One potential explanation for the weak relationship between explicit prejudice and implicit stereotyping involves the restricted range of responses on explicit prejudice measures. For example, the mean responses of participants in Experiments 2 and 3 on the Modern Racism Scale (*M*s = 1.66 and 1.89) were close to the absolute, nonprejudiced scale anchor of 1.00 for responses that could range from 1 to 5. The majority of respondents in Experiment 2 (84%) and Experiment 3 (62%) scored at or below the scale midpoint of 3, with standard deviations of .83 and .69. Nevertheless, restricted range is not the entire explanation for the weak relationship between these measures. Responses to the Attitudes Toward Blacks Scale (Brigham, 1993) had means closer to the scale midpoints in Experiments 2 and 3 (*M*s = 2.13 and 3.01), with more normal distributions and standard deviations of .50 and .62. In Experiment 1, responses on Swim, Aiken, Hall, and Hunter's (1995) Modern Sexism Scale, which could range from 1 to 7, ranged between 1.50 and 5.00. The mean on this scale ($M = 3.37$) also was close to the midpoint, with a standard deviation of .86.

REFERENCES

Anderson, S., & Klatzky, R. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology, 53*, 235-246.

Baker, S., & Devine, P. (1988, April). *Faces as primes for stereotype activation.* Paper presented at the 60th annual meeting of the Midwestern Psychological Association, Chicago.

Banaji, M., & Greenwald, A. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology, 68*, 181-198.

Banaji, M., & Hardin, C. (1996). Automatic stereotyping. *Psychological Science, 7*, 136-141.

Banaji, M., Hardin, C., & Rothman, A. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology, 65*, 272-281.

Bargh, J. (1994). The four horsemen of automaticity: Awareness, intention, efficiency and control in social cognition. In R. Wyer, Jr., & T. Srull (Eds.), *The handbook of social cognition* (2nd ed., pp. 1-40). Hillsdale, NJ: Lawrence Erlbaum.

Bargh, J., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of awareness on impression formation. *Journal of Personality and Social Psychology, 43*, 437-449.

Blair, I. (1999). Implicit stereotypes and prejudice. In G. Moskowitz (Ed.), *Cognitive social psychology: On the tenure and future of social cognition.* Mahwah, NJ: Erlbaum.

Blair, I., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142-1163.

Brauer, M., Wasel, W., & Niedenthal, P. (1999). *Implicit and explicit prejudice against women: The interrelationship among various components of sexism.* Manuscript submitted for publication.

Brewer, M. (1988). A dual process model of impression formation. In T. Srull & R. Wyer (Eds.), *Advances in social cognition: Vol. 1: A dual process model of impression formation* (pp. 1-36). Hillsdale, NJ: Lawrence Erlbaum.

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

Chen, M., & Bargh, J. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology, 33*, 541-560.

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

Dijksterhuis, A., & van Knippenberg, A. (1996). The knife that cuts both ways: Facilitated and inhibited access to traits as a result of stereotype-activation. *Journal of Experimental Social Psychology, 32*, 271-288.

Dovidio, J., Brigham, J., Johnson, B., & Gaertner, S. (1996). Stereotyping, prejudice, and discrimination: Another look. In N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 1276-1319). New York: Guilford.

Dovidio, J., Evans, N., & Tyler, R. (1986). Racial stereotypes: The contents of their cognitive representation. *Journal of Experimental Social Psychology, 22*, 22-37.

Dovidio, J., & Fazio, R. (1991). New technologies for the direct and indirect assessment of attitudes. In J. Tanur (Ed.), *Questions about survey questions: Meaning, memory, attitudes and social interaction* (pp. 204-237). New York: Russell Sage.

Dovidio, J., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33*, 510-540.

Fazio, R., Jackson, J., Dunton, B., & Williams, C. (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., Sanbonmatsu, D., Powell, M., & Kardes, F. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229-238.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Fiske, S. (1998). Stereotyping, prejudice, and discrimination. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 357-411). New York: McGraw-Hill.

Fiske, S., & Neuberg, S. (1990). A continuum of impression formation from category-based to individuating processes. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1-73). San Diego, CA: Academic Press.

Fiske, S., Neuberg, S., Beattie, A., & Milberg, S. (1987). Category-based and attribute-based reactions to others: Some informational conditions of stereotyping and individuating processes. *Journal of Experimental Social Psychology, 23*, 399-427.

Gaertner, S., & McLaughlin, J. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly, 46*, 3-30.

Glick, P., & Fiske, S. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*, 491-512.

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

Hense, R., Penner, L., & Nelson, D. (1995). Implicit memory for age stereotypes. *Social Cognition, 13*, 399-415.

Karlins, M., Coffman, T., & Walters, G. (1969). On the fading social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology, 28*, 280-290.

Katz, D., & Braly, K. (1933). Racial stereotypes in one hundred college students. *Journal of Abnormal and Social Psychology, 28*, 280-290.

Kawakami, K., Dion, K. L., & Dovidio, J. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin, 24*, 407-416.

Kawakami, K., Dovidio, J., Moll, J., Hermsen, S., & Russin, A. (1999). Just say no (to stereotyping): Effects of training in negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871-888.

Kihlstrom, J. (1990). The psychological unconscious. In L. Pervin (Ed.), *Handbook of personality theory and research: Theory and research* (pp. 445-464). New York: Guilford.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality and Social Psychology, 72*, 275-287.

Levin, D. (1996). Classifying faces by race: The structure of facial categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1364-1382.

Loftus, E., & Klinger, M. (1992). Is the unconscious smart or dumb? *American Psychologist, 47*, 761-765.

Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49*, 141-145.

Macrae, C., Bodenhausen, G., & Milne, A. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology, 69*, 397-407.

Macrae, C., Stangor, C., & Milne, A. (1994). Activating social stereotypes: A functional analysis. *Journal of Experimental Social Psychology, 30*, 370-389.

McClelland, D., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96*, 690-702.

McConahay, J. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. Dovidio & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-125). Orlando, FL: Academic Press.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Mullen, B. (1991). Group composition, salience, and cognitive representations: The phenomenology of being in a group. *Journal of Experimental Social Psychology, 27*, 297-323.

Mullen, B., Rozell, D., & Johnson, C. (1996). The phenomenology of being in a group: Complexity approaches to operationalizing cognitive representation. In J. Nye & A. Brower (Eds.), *What's social about social cognition? Research on socially shared cognition in small groups* (pp. 205-229). Thousand Oaks, CA: Sage.

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

Neely, J. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226-254.

Neely, J. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Lawrence Erlbaum.

Perdue, C., Dovidio, J., Gurtman, M., & Tyler, R. (1990). Us and them: Social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology, 59*, 475-486.

Perdue, C., & Gurtman, M. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology, 28*, 199-216.

Ratcliff, R. (1993). Methods for dealing with response time outliers. *Psychological Bulletin, 114*, 510-532.

Richardson-Klavehn, A., & Bjork, R. (1988). Measures of memory. *Annual Review of Psychology, 39*, 475-543.

Robinson, J., Shaver, P., & Wrightsman, L. (Eds.). (1991). *Measures of personality and social psychological attitudes.* San Diego, CA: Academic Press.

Salmon, D. P., & Butters, N. (1995). Neurobiology of skill and habit learning. *Current Opinion in Neurobiology, 5*, 184-190.

Schacter, D. (1990). Introduction to "Implicit memory: Multiple perspectives." *Bulletin of the Psychonomic Society, 28*, 338-340.

Spears, R., Oakes, P., Ellemers, N., & Haslam, A. (1997). *The social psychology of stereotyping and group life.* Oxford, UK: Blackwell.

Squire, L. R., & Kandel, E. R. (1999). *Memory: From mind to molecules.* New York: Freeman.

Stroessner, S. (1996). Social categorization by race or sex: Effects of perceived non-normalcy on response times. *Social Cognition, 14*, 247-276.

Swim, J., Aiken, K., Hall, W., & Hunter, B. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology, 68*, 199-214.

von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1997). The linguistic intergroup bias as an implicit indicator of prejudice. *Journal of Experimental Social Psychology, 33*, 490-509.

Wittenbrink, B., Judd, C., & 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.

Zárate, M., & Sandoval, P. (1995). The effects of contextual cues on making occupational and gender categorizations. *British Journal of Social Psychology, 34*, 353-362.

Zárate, M., & Smith, E. (1990). Person categorization and stereotyping. *Social Cognition, 8*, 161-185.

Zebrowitz, L., Montepare, J., & Lee, H. (1993). They don't all look alike: Individuated impressions of other racial groups. *Journal of Personality and Social Psychology, 65*, 85-101.

Received August 23, 1999
Revision accepted November 18, 1999