Research Report

RELATIONS BETWEEN IMPLICIT MEASURES OF PREJUDICE: What Are We Measuring?

Michael A. Olson and Russell H. Fazio

Ohio State University

Abstract—*Some recent findings suggest that different implicit measures of prejudice assess the same underlying construct, but other work suggests that they may not. In this experiment, White participants completed a version of a priming measure of racial attitudes that either encouraged categorization of the face primes in terms of race or did not encourage such categorization, and then completed the Implicit Association Test. Correspondence between the two measures was found only when categorization by race was required on the priming measure. Moreover, participants appeared more prejudiced when they were led to construe individuals in terms of race than when they were not so encouraged. The discussion focuses on the potential for dissociations between evaluations of a category and evaluations of members of the category.*

Asking someone to report his or her attitude toward another race may not produce an honest response. Implicit measures do not require respondents to report an attitude and are less controllable by respondents, so they appear to solve the social-desirability biases of explicit measures (Fazio & Olson, 2003). The present research addresses the correspondence between two implicit measures of attitudes: the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and a priming technique sometimes referred to as the "bona fide pipeline" (BFP; Fazio, Jackson, Dunton, & Williams, 1995).

The IAT measures the associative strength between two target categories (e.g., Blacks and Whites) and two attributes (e.g., pleasant and unpleasant) by forcing participants to categorize exemplars of both the target and the attribute categories within a single task. Negativity toward Blacks is evident in faster response latencies on Black-unpleasant (and White-pleasant) trials than on Black-pleasant (and White-unpleasant) trials. The BFP assesses the evaluation activated in response to a prime by considering how the prime (e.g., a Black or White face) facilitates judging the connotation of subsequently presented evaluative adjectives. Prejudice toward Blacks is evident in faster latencies to negative adjectives (and slower latencies to positive adjectives) following Black compared with White primes.

Both measures have been shown to predict race-related behaviors (Fazio & Olson, 2003), and several researchers have argued that, apart from measurement error and procedural differences, they should correspond to one another (e.g., Banaji, 2001). In fact, Cunningham, Preacher, and Banaji (2001) found that correspondence between the measures improved from around .20 to over .50 after latent structural equation modeling was used to control for low reliabilities.

However, evidence suggests that the IAT and BFP may measure different constructs. In our own lab, four studies with more than 300 participants altogether have revealed little correspondence between them (*r*s from .05 to −.13). Correlations of essentially zero also have been reported for smoking attitudes (Sherman, Presson, Chassin, Rose, & Koch, 2003) and condom use (Marsh, Johnson, & Scott-Sheldon, 2001). Although measurement error undoubtedly plays a role, it probably cannot fully account for such null relations.

Another difference between the measures is the percentage of participants who appear prejudiced on each. The BFP reveals negativity in 50 to 60% of White college students (e.g., Fazio et al., 1995), but prejudiced IAT scores are found in 70 to 90% of Whites (e.g., Nosek, Banaji, & Greenwald, 2002).

That the two measures correlate sporadically at best and show different distributions of prejudice implies some difference in the psychological constructs they tap. Consideration of the mechanism underlying each measure may point to the nature of that difference and provide insight into one condition in which they might correspond (see Fazio & Olson, 2003, for a detailed analysis). In the BFP, positivity or negativity is automatically activated in response to an attitude-evoking prime, which readies an evaluatively congruent response. Evaluatively congruent adjectives are responded to relatively quickly, and response competition slows responses to incongruent adjectives (see Fazio, 2001, for a review). The BFP typically includes exemplars of two categories as primes (e.g., Black and White faces), and responses are averaged across exemplars to estimate attitudes toward the categories. It is important to note that responses are made at the level of the individual exemplar, and participants are not forced to construe primes as members of a particular category. This sensitivity to specific exemplar primes was illustrated by Livingston and Brewer (2002), who observed greater automatically activated negativity in response to prototypical compared with less prototypical Black faces. This difference, however, was eliminated when participants were instructed to attend to race.

The IAT is based on the assumption that two categories that are associated in memory (e.g., Blacks and unpleasant) will be more easily represented by the same response key (Greenwald & Nosek, 2001) than two categories that are not associated. De Houwer (2001) suggested that associations to categories drive the IAT more than do specific exemplars. In a British-foreigner IAT that included both liked and disliked Brits (princess Diana, a mass murderer) and foreigners (Einstein, Hitler), British participants showed a bias toward Brits regardless of the valence of the specific exemplars (De Houwer, 2001). This suggests that the IAT is affected more by associations to category labels than by evaluations activated by a given exemplar (see also Mitchell, Nosek, & Banaji, in press).

The BFP bases scores only on the evaluation automatically activated in response to an exemplar, which may or may not include category-level information. This implies that forcing participants to construe the exemplar primes as representatives of the category, as in Livingston and Brewer's (2002) experiment, will produce responses that tap associations to the category, resulting in increased BFP-IAT

Address correspondence to Michael Olson, Department of Psychology, 1885 Neil Ave., Ohio State University, Columbus, OH 43210-1222; e-mail: olson@psy.ohio-state.edu.

correspondence. The experiment reported here was designed to test this hypothesis.

The BFP involves a cover story justifying the presence of the primes. Participants are told that if judging word meaning occurs automatically, then they should be able to perform the adjective-connotation task well, even while performing another task simultaneously. We used this secondary task to manipulate whether participants were free to construe the faces as they would normally, as in the "traditional" BFP, or were forced to categorize them by race, as in the "category" BFP. We predicted better correspondence between the IAT and the category BFP than between the IAT and the traditional BFP.

METHOD

One hundred White undergraduates participated for course credit. Participants with high error rates ($>$ 20%) on either measure were omitted, resulting in a sample of 61 females and 31 males. They were told that they would be participating in two separate experiments, the first (BFP) about word meaning as an automatic skill and the second (IAT) about categorization skills.

The BFP procedure (for more details, see Fazio et al., 1995) involved multiple phases. In Phase 1, participants identified the connotation of 12 positive and 12 negative adjectives by pressing either a "good" or a "bad" key. In Phase 2, Black, White, Asian, and Latino faces were presented. In the traditional condition, participants were told, "We're interested in how well you can learn these faces, so it's important that you pay attention to them. After you finish this task, we are going to test you for how well you can recognize these faces." In the category condition, they were told, "We want you to keep a mental tally of how many of the faces were Caucasian, Asian, Latino, and African-American. After you finish this task, we are going to have you estimate how many members of each race you saw." Phase 3 consisted of the test that participants anticipated.

Participants were told that Phase 4 (the priming phase) combined Phases 1 and 2, and consisted of four blocks. On a given trial, a prime, which participants were to either study or add to their racial tally, was presented for 315 ms, followed by a 135-ms interval and then the target adjective. Participants responded to the target as in Phase 1. Thirty-two of the 48 trials per block included a prime from 16 gender-matched Black-White pairs presented with the same two positive and two negative adjectives. Primes were yearbook-style color photos (and included other-race fillers). Participants in the category condition estimated the number of faces presented for one of the four races after each block. In the traditional condition, participants completed a face recognition test at the end of the priming phase. They were then escorted to another area of the lab.

The IAT included 12 blocks of 50 trials each. On a given trial, participants were presented with an exemplar of one of four categories: Black names, White names, pleasant words, and unpleasant words (stimuli were from Greenwald et al., 1998). Participants categorized items by pressing one of two keys whose meanings changed depending on the block. Participants categorized Black and White names in Blocks 1 and 2, and pleasant and unpleasant items in Blocks 3 and 4. Blocks 5 through 7 were critical combined blocks, in which one of the races and pleasant words were assigned to one response key, and the other race and unpleasant words were assigned to the other response key (counterbalanced). Blocks 8 and 9 involved categorizing Black and White names, with the meaning of the keys now reversed. Blocks 10 through 12 were identical to Blocks 5 through 7, but the race that was associated with pleasant items was now associated with unpleasant items (and vice versa).

Table 1. *Descriptive data for each implicit measure*

Measure	Mean	*SD*	Proportion prejudiced
Traditional BFP	0.00	.26	.52
Category BFP	−.19	.33	.74
IAT	79.6 ms	91.5	.79

Note. For the bona fide pipeline (BFP), more positive numbers reflect more positivity toward Blacks; the reverse is true for the Implicit Association Test (IAT). The last column refers to the proportion of participants with scores on the side of the neutral point indicative of prejudice toward Blacks.

RESULTS

BFP

Attitude estimates were derived as described in Fazio et al. (1995). For each participant, mean facilitation scores for the two positive and two negative adjectives were computed for each face. An effect size of the Race of Prime $\times$ Valence of Adjective interaction was computed for each participant, resulting in an attitude estimate in which negative numbers imply more negativity toward Blacks than Whites (see Table 1). Participants' scores were more negative in the category condition than in the traditional condition, $t(90) = 3.06$, $p < .01$, with the former mean differing significantly from zero, $t(42) = 3.67$, $p < .01$.[1]

IAT

IAT scores were computed as described in Greenwald et al. (1998). The first two trials from each block were dropped, and response latencies were natural-log-transformed. The mean from the three blocks involving White-pleasant and Black-unpleasant pairings was subtracted from the mean from the blocks involving White-unpleasant and Black-pleasant pairings, resulting in a measure for which higher numbers indicate more negativity toward Blacks (see Table 1). On average, participants appeared prejudiced against Blacks, $t(91) = 9.12$, $p < .001$. IAT scores did not vary as a function of BFP condition, $t < 1$.

Proportion Appearing Prejudiced

The proportion of participants displaying some degree of negativity toward Blacks (see Table 1) was significantly lower for the traditional BFP than for either the category BFP ($p < .05$) or the IAT ($p < .01$).

1. In the many studies we have conducted using the BFP, the average score has sometimes been significantly more negative than zero (e.g., Fazio et al., 1995; Olson & Fazio, 1999; Towles-Schwen & Fazio, 2001) and sometimes not (e.g., Fazio & Dunton, 1997; Fazio & Hilden, 2001; Jackson, 1997; Olson & Fazio, in press; Towles-Schwen, 2002). We presume this simply reflects sampling variability. Relations between the attitude estimates and race-related judgments and behaviors have been observed regardless of the sample's average negativity toward Blacks.

BFP-IAT Correspondence

A regression analysis predicting IAT scores from BFP scores, a condition dummy variable, and the interaction term revealed a significant BFP Score × Condition interaction, $t(87) = 2.07$, $p < .05$. The category BFP corresponded with the IAT, $\beta = -.28$, $t(40) = 2.03$, $p = .04$, but the traditional BFP did not, $\beta = .18$, $t < 1$.[2]

Reliability

Split-half correlation coefficients were computed using attitude estimates based on the first and second halves of the critical trials for each measure. Correlations were .04 (n.s.) and .39 ($p < .05$) for the traditional and category BFP, respectively, and .53 ($p < .05$) for the IAT.

DISCUSSION

Confirming our reasoning that they measure different constructs, a traditional version of the BFP and the IAT showed little correspondence. However, correspondence was observed when participants were forced to categorize exemplars as representatives of racial categories during the BFP. These results are consistent with our reasoning that the BFP assesses evaluations of exemplars and the IAT assesses associations to categories.[3] We reconcile these findings with those of Cunningham et al. (2001) by noting that their participants completed several explicit measures of prejudice, that their priming procedure used only Black and White faces, and that it sometimes was completed after the IAT. Hence, their procedures made race salient, encouraging categorization by race, much as the category version of the BFP does.

The distribution of prejudice also showed an interesting pattern. Roughly three quarters of the participants appeared prejudiced on the IAT and the category version of the BFP, compared with about half on the traditional BFP. Thus, it appears that evaluations of Blacks are more negative when assessed at the category level than when assessed at the level of the exemplar, a finding that extends Sears's (1983) notion of more favorable self-reported evaluations of exemplars than collectives to implicit measures. Although it may appear surprising that evaluations of a category can be somewhat distinct from evaluations of the category exemplars, the informational environment might encourage such dissociations. For example, "Blacks" are often represented negatively without reference to individual members, and individual Black celebrities are often represented positively without reference to their category membership.

We argued that the category BFP related to the IAT more strongly than the traditional BFP did because both the category BFP and the IAT assessed category-level associations. In our view, the observed difference in reliability between the two versions of the BFP also reflects their differential emphasis on exemplar- versus category-level construal. In the traditional BFP, people are free to construe the faces as they do naturally; they need not categorize by race (Fazio & Dunton, 1997). They may, for example, attend to the gender of some faces, and to the attractiveness of others. Thus, for people who do not spontaneously attend to race, the estimate of racial attitudes will be essentially noise, because it is based on Black-White difference scores. People with more extreme racial-attitude estimates, in contrast, are known to categorize social targets by race more extensively (Fazio & Dunton, 1997). They also displayed more reliability on the traditional BFP in the current study.[4] So what appears to be poor reliability based on simple measurement error is at least in part based on real differences regarding spontaneous categorization by race. Because it forces categorization by race, the category BFP provides both a reliable estimate of reactions to the Black versus White faces and correspondence with the IAT.

It is important to note that the traditional BFP has proven to be a reliable predictor of behavior in past studies (see Fazio & Olson, 2003). Given the many demonstrations of the predictive validity of the traditional BFP, it seems inappropriate to dismiss the lack of a relation between the BFP and the IAT as due to the former's unreliability. Although allowing categorization by race to vary reduces the traditional BFP's reliability, that same natural variation may make it a relatively superior predictor of judgments and behavior toward individual Blacks in settings that do not promote categorization by race. In contrast, behavior toward the category "Black," or toward an individual Black in settings that do encourage categorization by race, may be better predicted by the IAT or the category version of the BFP.

Acknowledgments—This research was supported by National Institute of Mental Health Grant MH38832 and Senior Scientist Award MH01646. Julie Brown, Eric Currence, Nicole Deland, Don Ladwig, Sarah Lifland, and Suzanne Miller assisted in data collection, and Marilynn Brewer provided helpful feedback.

REFERENCES

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

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. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*, 443–451.

Fazio, R.H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*, 115–141.

Fazio, R.H., & Dunton, B.C. (1997). Categorization by race: The impact of automatic and controlled components of racial prejudice. *Journal of Experimental Social Psychology*, *33*, 451–470.

Fazio, R.H., & Hilden, L.E. (2001). Emotional reactions to a seemingly prejudiced response: The role of automatically-activated racial attitudes and motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *27*, 538–549.

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., McGhee, D., & Schwartz, J.L.K. (1998). Measuring individual differences in cognition: The implicit association task. *Journal of Personality and Social Psychology*, *74*, 1469–1480.

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

Jackson, J.R. (1997). *Automatically activated racial attitudes.* Unpublished doctoral dissertation, Indiana University, Bloomington.

2. IAT block was also included and, as is typical, accounted for significant variance, $t(87) = 1.97$, $p = .05$.

3. The BFP and IAT may not correspond for other reasons as well (see, e.g., Karpinski & Hilton, 2001, and Fazio & Olson, in press, for consideration of the potential effects of environmental associations that are discrepant from personal evaluations).

4. For the 15 participants who showed the largest race differences (positive or negative) in their responses on the traditional BFP, split-half reliability improved to .46. The comparable figure for the category BFP was .49.

Karpinski, A., & Hilton, J.L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 774–778.

Livingston, R.W., & Brewer, M.B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, *82*, 5–18.

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

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

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. (1999, May). *Nonverbal leakage during public evaluations of Black candidates: The roles of automatically-activated racial attitudes and motivation to control prejudiced reactions.* Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago.

Olson, M.A., & Fazio, R.H. (in press). Trait inferences as a function of automatically-activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology.*

Sears, D.O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, *44*, 233–250.

Sherman, S.J., Presson, C.C., Chassin, L., Rose, J.S., & Koch, K. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*, 13–39.

Towles-Schwen, T. (2002). *White students' relationships with their African-American roommates: Automatically-activated racial attitudes and motivation to control prejudiced reactions as antecedents and consequences.* Unpublished doctoral dissertation, Indiana University, Bloomington.

Towles-Schwen, T., & Fazio, R.H. (2001). On the origins of racial attitudes: Correlates of childhood experiences. *Personality and Social Psychology Bulletin*, *27*, 162–175.

(RECEIVED 7/23/02; REVISION ACCEPTED 2/4/03)