Original Article

Changing Implicit and Explicit Prejudice

Insights from the Associative-Propositional Evaluation Model

Rajees Sritharan and Bertram Gawronski

The University of Western Ontario, London, Ontario, Canada

Abstract. Although overt prejudice has declined in many societies over the past decades, new advancements in intergroup relations research have uncovered various kinds of subtle biases that continue to prevail despite increases in egalitarian values. Understanding the processes that may produce inconsistencies between spontaneous affective responses and self-reported explicit evaluations can provide deeper insights into conceptually different forms of prejudice, including both overt and subtle variants. In the present article, research on prejudice reduction is reviewed from the perspective of the associative-propositional evaluation (APE) model, which considers evaluations through the processes of associative activation and propositional validation. The APE model's potential for integrating different conceptualizations of overt and subtle prejudice and the application of the model to prejudice reduction are discussed.

Keywords: attitude change, explicit evaluation, implicit evaluation, prejudice, stereotypes

Common mantras among individuals who do not wish to believe themselves to be prejudiced are qualifying phrases like "I'm not a racist, but . . ." before expressing a racial stereotype such as, "Black people are lazy" (Bonilla-Silva & Forman, 2000). Although such overt expressions of prejudice have become less socially acceptable over time (Sniderman & Carmines, 1997), there has been surprisingly little concurrent reduction in intergroup conflicts, which have only dropped slightly over recent decades (Dovidio & Gaertner, 2004). Furthermore, there has been some strong resistance to policies attempting to reduce intergroup divisions, such as affirmative action and desegregation of schools, even among individuals believing in egalitarian ideals (Jacobson, 1985). Similarly, whereas support for legislation prohibiting interracial marriage has markedly declined over time, the actual rate of interracial marriage has remained relatively low and stable over this same period (Fryer, 2007).

These apparent disconnects between stated beliefs and actual behavior have been interpreted as prejudice becoming expressed in more subtle and indirect forms (Crosby, Bromley, & Saxe, 1980) since old-fashioned prejudice, characterized by overt expressions of negativity and open support for segregation and discrimination, has become more socially unacceptable over time (McConahay, Hardee, & Batts, 1981). Intergroup relations theorists have attempted to understand the changing face of prejudice by examining these more subtle manifestations of prejudice from perspectives including symbolic racism (Sears & Henry, 2003), modern racism (McConahay, 1986), aversive racism (Dovidio & Gaertner, 2004; Gaertner & Dovidio, 1986), ambivalent racism (Katz & Hass, 1988), automatic prejudice (Devine, 1989), and implicit prejudice (Rudman, Greenwald, Mellott, & Schwartz, 1999).

The continued presence of subtle forms of prejudice has also been supported by data obtained from a new class of indirect measurement procedures (for a review, see Fazio & Olson, 2003). Researchers studying attitudes have traditionally examined only explicit evaluations, which are typically described as consciously reported, deliberate evaluations. More recently, the focus has shifted toward the inclusion of implicit evaluations, which can be described as spontaneous evaluative responses that are unintentional, difficult to control, but not necessarily unconscious (Gawronski, Hofmann, & Wilbur, 2006). Whereas explicit evaluations are often equated with deliberate judgments reflected in self-report measures, implicit evaluations are inferred from task performance on indirect measures, such as the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) or different kinds of sequential priming tasks (e.g., Fazio, Jackson, Dunton, & Williams, 1995; Payne, Cheng, Govorun, & Stewart, 2005; Wittenbrink, Judd, & Park, 1997).

DOI: 10.1027/1864-9335/a000017

Interest in using indirect measures to study prejudice began in the 1980s, generating a multitude of measurement procedures and data in the ensuing years. In one of the earliest studies to make use of indirect measures for this purpose, Gaertner and McLaughlin (1983) found that participants showed quicker responses in a sequential priming task when the prime-target pairs involved stereotypical combinations of racial groups and traits than when the prime-target pairs were nonstereotypical. Since Gaertner and McLaughlin's seminal work, a substantial number of studies found that explicit evaluations show better performance in predicting deliberate behavior, whereas implicit evaluations are better predictors of spontaneous behavior (for a review, see Friese, Hofmann, & Schmitt, 2008). These results indicate the importance of considering both overt and subtle variants of prejudice and the processes which underlie them in order to develop effective prejudice reduction strategies.

The present article illustrates the usefulness of the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006a, 2006b, 2007) in providing deeper insights into (1) the processes underlying overt and subtle forms of prejudice, and (2) the relative effectiveness of different strategies in changing prejudice. For this purpose, we first provide a brief overview of earlier accounts of implicit and explicit evaluations, and how they interpret overt and subtle forms of prejudice. Expanding on these models, we then outline the core assumptions of the APE model and how these assumptions can be applied to integrate different types of overt and subtle prejudice. In the remainder of the article, we discuss the implications of this integration for prejudice reduction, reviewing relevant research with a particular focus on the mechanisms underlying different types of evaluative responses. Based on this review, we conclude by outlining a number of directions for future research on prejudice reduction.

Earlier Accounts of Implicit and Explicit Evaluations

The MODE Model

The MODE model defines attitude as the association between an object and an evaluation in memory, which may vary in strength (Fazio, 2007; Olson & Fazio, 2009). A strong object-evaluation association may result in this evaluation being automatically activated when encountering the object, without an individual necessarily having any intent to evaluate the object. According to this model, the key determinants of whether automatically activated attitudes are used in self-reported evaluations are motivation and opportunity to engage in effortful processing. According to the MODE model, automatic attitudes will be the primary determinant of downstream evaluations when either motivation or opportunity is low. If, however, motivation and opportunity are high, the impact of automatic attitudes on self-reported evaluations may be diluted or inhibited by deliberate processes.

With regard to prejudice, the MODE model argues that motivation to control prejudiced reactions is a crucial factor in the expression of prejudice (e.g., Fazio et al., 1995). The MODE model further asserts that indirect measures are a good proxy for automatic attitudes because they tend to provide little opportunity to control automatically activated evaluations. As such, individuals who have high motivation to control prejudice may display low levels of prejudice on self-report measures, but may nevertheless show negative automatic evaluations of an outgroup on indirect measures that do not provide the opportunity to control automatic reactions. Thus, the MODE model explains dissociations between implicit and explicit prejudice as the product of motivation and opportunity to control the impact of automatic negative attitudes on self-reported evaluative judgments. However, the MODE model does not further distinguish between other variants of prejudice, such as old-fashioned, modern, and aversive forms.

The Dual Attitudes Model

Wilson, Lindsey, and Schooler's (2000) dual attitudes model assumes that the same object can be represented with two distinct evaluations in memory, each requiring different levels of effort to retrieve. This model suggests that earlier acquired attitudes are highly stable and activated automatically. Moreover, recently acquired attitudes are assumed to coexist with older attitudes, though their retrieval typically requires more cognitive effort compared to the automatic activation of earlier acquired attitudes. According to the dual attitudes model, the extent to which self-reported evaluations reflect earlier versus recently acquired attitudes depends on the motivation and the cognitive capacity to engage in the effortful process of retrieving the new attitude from memory.

Applied to the subject of prejudice, the dual attitudes model may consider implicit prejudice to be the result of early socialization experiences, and this form of prejudice may be particularly resistant to change. However, individuals may acquire more positive attitudes later in the socialization process, which may be expressed in self-report measures when they engage in the effortful processing necessary to retrieve these attitudes from memory. As such, the dual attitudes model considers dissociations between implicit and explicit prejudice as the result of early versus recent learning experiences and the cognitive effort required to retrieve recently acquired, nonprejudiced attitudes from memory. However, as with the MODE model, the dual attitudes model does not further distinguish between other variants of prejudice, such as old-fashioned, modern, and aversive forms. Furthermore, the dual attitudes model's implication that explicit prejudice should be more context-sensitive than implicit prejudice stands in

contrast to research demonstrating that implicit prejudice can in fact be more responsive to contextual variations than explicit prejudice (e.g., Dasgupta & Greenwald, 2001).

Associative and Propositional Processes in Prejudice

The APE Model

The APE model (Gawronski & Bodenhausen, 2006a, 2006b, 2007) posits that evaluations of social and nonsocial objects should be understood by focusing on their underlying mental processes, namely associative and propositional processes (see also Strack & Deutsch, 2004). Associative processes involve the *activation* of evaluative associations in reaction to a stimulus, which in turn determine an individual's spontaneous affective response (*implicit evaluation*). Propositional processes entail the *validation* of the evaluation implied by the affective response. That is, through propositional reasoning, individuals determine whether or not their affective response provides a valid basis for a deliberate judgment that is explicitly endorsed (*explicit evaluation*). Thus, affective responses resulting from spontaneously activated associations may be accepted or rejected through processes of propositional validation.

Because associative processes are assumed to operate independently of validation, spontaneous affective responses may be activated irrespective of whether or not an individual deliberately endorses the evaluation implied by the affective response. Thus, associative activation and propositional validation can lead to conflicting evaluative responses to the same stimulus. Such dissociations occur when one's affective response is rejected propositionally as a valid source for making an evaluative judgment. According to the APE model, the acceptance or rejection of an affective response depends on the consistency of this response with other momentarily considered propositional information. If the affective response is consistent with this information, it can be accepted, and therefore be used as a basis for an evaluative judgment. If, however, the affective response is inconsistent with other momentarily considered propositions, it may be rejected as a basis for an evaluative judgment (e.g., Gawronski & Strack, 2004; Gawronski, Peters, Brochu, & Strack, 2008). For example, if one experiences a spontaneous negative reaction when encountering a member of a stigmatized group, one may accept the affective response and therefore use it as a basis for the endorsed propositional evaluation *I dislike this person.* However, the evaluation implied by the negative affective response may also be rejected through propositional invalidation because of its inconsistency with other propositional beliefs. For instance, if one endorses the propositional belief that *This person belongs to a stigmatized group* and also believes that *Treating individuals belonging to stigmatized groups negatively is wrong,* one may be compelled to invalidate the negative affective response and instead accept the propositional evaluation *I like this person,* or endorse a neutral evaluation. In such cases, invalidation of the affective response leads to dissociations between spontaneous affective responses and endorsed evaluative judgments, which can be reflected in divergent scores on indirect and self-report measures of prejudice (e.g., Dunton & Fazio, 1997; Franco & Maass, 1999; for a review, see Hofmann, Gschwendner, Nosek, & Schmitt, 2005).

Integrating Different Types of Prejudice

The proposed role of cognitive consistency in modulating the relation between affective responses and endorsed evaluations has recently been applied by Gawronski, Peters et al. (2008) to integrate conceptually different forms of prejudice. In their unifying framework, several distinct types of prejudice that have been discussed in the literature (e.g., old-fashioned prejudice, modern prejudice, aversive prejudice, implicit prejudice) can be integrated by means of their constituent evaluative processes. The central components in this integrative framework are: (a) spontaneous affective reactions, (b) the propositional evaluation implied by the affective reaction, (c) egalitarianism-related, nonprejudicial goals, and (d) perceptions of discrimination. The proposed interactions between these components are illustrated for racial attitudes in Figure 1. Panel A depicts an inconsistent belief system that results from the acceptance of a negative affective response for a propositional evaluative judgment (*I dislike Black people*) despite the endorsement of nonprejudicial goals (*Negative evaluations of disadvantaged minority groups are wrong*) and high levels of perceived discrimination (*Black people represent a disadvantaged minority group*). To avoid cognitive dissonance (Festinger, 1957), this inconsistency could be resolved by (a) rejecting nonprejudicial goals (Panel B), (b) denying continued discrimination (Panel C), or (c) rejecting the negative affective response as a basis for a propositional judgment (Panel D). Whereas the pattern depicted in the first case – rejection of nonprejudicial goals – reflects the central notion of old-fashioned prejudice (Swim, Aikin, Hall, & Hunter, 1995), the second case – denial of continued discrimination – has become a central tenet in theories of modern prejudice (McConahay, 1986). Finally, the third case – rejection of negative affective responses – can be understood as the type of ambivalence discussed by theories of aversive racism, which has been conceptualized as the conflict between negative feelings and strong egalitarian beliefs (Dovidio & Gaertner, 2004).

The proposed framework not only provides a theoretical integration of different kinds of overt and subtle prejudices, it also implies a number of predictions about the relation between implicit and explicit prejudice, as assessed by indirect and self-report measures. Specifically, the proposed framework implies that spontaneous affective reactions (implicit prejudice) should correspond with self-reported

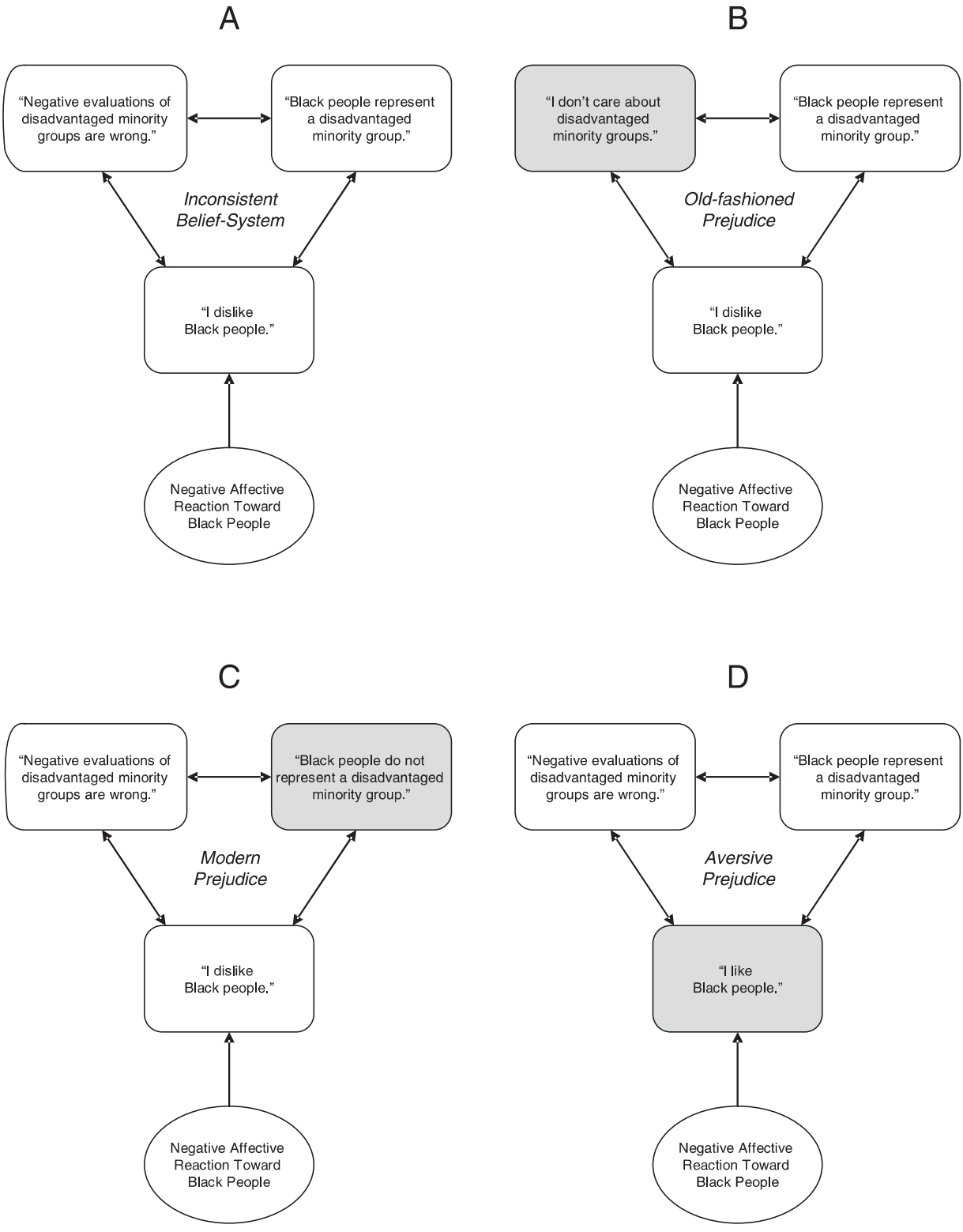

Figure 1. Interplay between spontaneous affective reactions (circles) and propositional beliefs (squares) in racial prejudice against Black people. Panel A depicts the case of an inconsistent belief system resulting from negative affective reactions toward Black people; Panels B, C, and D depict consistent belief systems as they are reflected in old-fashioned, modern, and aversive prejudice, respectively. Figure adapted from Gawronski, Peters, Brochu, & Strack, 2008a; reprinted with permission.

evaluative judgments (explicit prejudice) if either nonprejudicial goals or perceptions of discrimination are low (Panels B and C). However, the two kinds of evaluative responses should be dissociated if both nonprejudicial goals and perceptions of discrimination are high (Panel D). These predictions have been empirically confirmed by Gawronski, Peters et al. (2008) for racial prejudice and by Brochu, Gawronski, and Esses (2010) for prejudice against the overweight (for a review, see Brochu, Gawronski, & Esses, 2010).

Prejudice Reduction

In addition to providing a unifying framework for the integration of different forms of prejudice, the APE model also has a number of implications for changing prejudice. As may already be evident from Gawronski, Peters et al.'s (2008) integrative framework (see Figure 1), targeting single components of prejudice-related belief systems may often be ineffective in producing the desired changes in overt behavior. For instance, simply increasing the awareness of ongoing social discrimination may be unsuccessful in reducing prejudiced responses when nonprejudicial goals are weak (as depicted in Panel B). Moreover, successful attempts to enhance nonprejudicial goals may leave prejudiced responses unaffected when the target group is not considered to be discriminated against (as depicted in Panel C).[1] Thus, strategies that can be directed at all of the relevant components simultaneously may be most successful in producing the desired changes in overt prejudicial responses.

Changes in Propositional Beliefs

According to the APE model, changes in propositional beliefs can be described as a change in the subjective truth or falsity of a given proposition (Gawronski & Bodenhausen, 2006a). Such changes in the ascription of truth values are typically due to cognitive inconsistency resulting from a change in the set of considered propositions. One important factor in this context is persuasive communication. Exposing people to persuasive messages typically provides additional information to the current set of beliefs, which in turn may produce changes in propositional beliefs as result of the desire to maintain cognitive consistency (for a more detailed discussion, see Festinger, 1957, Chapter 6). In line with this reasoning, prejudice has often been targeted with persuasive campaigns that appeal to people's values and beliefs with the goal of questioning the morality or legality of prejudice (for a review, see Amodio & Devine, 2005). The general expectation is that by increasing the salience of prejudice-inconsistent values or beliefs, prejudiced individuals may change their propositional beliefs about the outgroup, following the general principles identified in persuasion research (for a review, see Petty, Cacioppo, Strathman, & Priester, 2005). Applied to Gawronski, Peters et al.'s (2008) integrative framework, these insights could be used to increase nonprejudicial goals (e.g., by appealing to personal values) and perceptions of discrimination (e.g., by providing factual information), which represent two of the central components of prejudice-related belief systems (see Figure 1).

Another important source of cognitive inconsistency is counterattitudinal behavior (Festinger & Carlsmith, 1959). In line with Festinger's (1957) theory of cognitive dissonance, Leippe and Eisenstadt (1994) found that White participants asked to write an essay supporting an increase in scholarships for Blacks reported less negative attitudes toward Blacks when the situational pressure to write the essay was low rather than when the situational pressure was high. Expanding on these findings, Gawronski and Strack (2004) asked non-Black students to write an essay in favor of a policy change that would double scholarships for Black students under either low or high situational pressure. Replicating Leippe and Eisenstadt's (1994) findings, Gawronski and Strack (2004) found lower levels of explicit prejudice under low situational pressure. However, implicit prejudice remained high regardless of situational pressure. Moreover, measures of explicit and implicit prejudice were positively correlated when situational pressure was high, but uncorrelated when situational pressure was low. According to Gawronski and Strack (2004), these results indicate that cognitive dissonance is an inherently propositional phenomenon, supporting a central assumption of the APE model (Gawronski & Bodenhausen, 2006a). In addition, Gawronski and Strack's (2004) findings suggest that dissonance manipulations could possibly be used to change propositional components of prejudice-related belief systems. However, they may be less suited to changing negative affective reactions resulting from automatic associations.

Changes in Affective Reactions

From the perspective of the APE model, spontaneous affective reactions are the product of evaluative associations that are activated automatically upon the encounter of an attitude object (Gawronski & Bodenhausen, 2007). Such automatic associations are typically assessed with indirect measures, such as the IAT (Greenwald et al., 1998) or var-

[1] In this context, it seems worth noting that the recent election of Barack Obama as the first African American president of the United States may in fact have paradoxical effects, in that his election may reduce perceptions of continued discrimination and therefore "allow" the propositional endorsement of negative affective reactions toward African Americans (see Figure 1, Panel C; for related results, see Effron, Cameron, & Monin, 2009; Kaiser, Drury, Spalding, Cheryan, & O'Brien, 2009).

ious kinds of sequential priming tasks (e.g., Fazio et al., 1995; Payne et al., 2005; Wittenbrink et al., 1997). The development of these measures has also stimulated research regarding the malleability of automatic associations. Although experimental paradigms have been shown to differ in their effectiveness in changing automatic associations (e.g., Gawronski & LeBel, 2008; Gawronski & Strack, 2004), significant changes have been shown to be possible to some extent (for reviews, see Blair, 2002; Gawronski & Bodenhausen, 2006a). According to the APE model, changes in spontaneous affective reactions can be due to (1) momentary changes in the activated pattern of associations or (2) changes in the underlying structure of chronic associations.

Changes caused by variations in pattern activation can occur when contextual cues activate different subsets of the associative representation of an object. For instance, responses to the same individual (e.g., Michael Jordan) may differ as a function of whether this person is categorized in terms of race or an alternative category, such as occupation (Mitchell, Nosek & Banaji, 2003). Consequently, evaluative responses are often context-sensitive, in that contextual cues may influence the categorization of a given individual (Fazio, 2007). Investigating reverse influences from exemplars to groups, Dasgupta and Greenwald (2001) found that exposing participants to pictures of admired Blacks and disliked Whites reduced implicit preferences for Whites over Blacks. These results suggest that momentary construals of social groups may differ as a function of highly accessible exemplars, which in turn influences spontaneous affective reactions to these groups (Smith & Zárate, 1992). However, in evaluating the effectiveness of these mechanisms in reducing prejudice, it should be noted that the resulting changes reflect only momentary variations in activated associations which can differ as a function of context and time (see Joy-Gaba & Nosek, 2010). As such, they may be less suited to bring about long-lasting and context-independent changes in overt behavior, which would be desirable in the context of prejudice and stereotyping.

A strategy that is more likely to produce stable outcomes is to change the chronic associations that are responsible for enduring affective responses. One procedure that may produce such changes is the evaluative conditioning (EC) paradigm, which repeatedly pairs positive or negative unconditioned stimuli (US) with neutral conditioned stimuli (CS) to produce changes in attitudes toward the CS (for a review, see De Houwer, Thomas, & Baeyens, 2001). In the domain of racial prejudice, Olson and Fazio (2006) employed an EC paradigm with Black and White faces as CS and positive and negative stimuli as the US. Following the EC procedure, implicit prejudice was measured with an affective priming task (Fazio et al., 1995), using new Black and White faces as primes. Their results showed reduced implicit preferences for Whites over Blacks when Black faces were repeatedly paired with positive stimuli and White faces were repeatedly paired with negative stimuli. Moreover, these effects remained stable when measured two days later. These changes in implicit preferences, however, did not produce corresponding changes on self-report measures of explicit prejudice. This dissociation between changes in implicit and explicit evaluations corroborates the concern that targeting single components in the overall belief system – in this case, automatic associations – may be less effective in producing generalized changes in overt behavior. Instead, reducing prejudice might require targeting all the components of prejudice-related belief systems, not just automatic associations that may or may not be accepted as valid basis for explicit evaluative judgments (Gawronski & LeBel, 2008).

As previously discussed, explicit evaluations can be changed with counterattitudinal information, and these changes likely follow the principles identified in research on persuasion (Petty et al., 2005). However, in line with the notion that a given manipulation may be differentially effective in changing explicit and implicit evaluations (see Gawronski & Bodenhausen, 2006a), Rydell, McConnell, Strain, Claypool, and Hugenberg (2007) determined that greater amounts of counterattitudinal information changed implicit evaluations in a slow and fairly linear manner, whereas explicit evaluations changed rapidly in response to counterattitudinal information. From the perspective of the APE model, one could argue that the counterattitudinal information quickly invalidates prior explicit evaluations, whereas changes in implicit evaluations depend on the incremental integration of counterattitudinal information into the associative representation of the object. Examining evaluative generalizations from exemplars to groups, Ranganath and Nosek (2008) found that these processes can also produce the opposite pattern when they involve the formation rather than the change of associative representations. In their study, newly acquired information about an individual quickly generalized to the individual's unfamiliar group at the implicit level, whereas generalization at the explicit level occurred only after several days. Using the APE model, Ranganath and Nosek (2008) interpreted these findings as indicating that simple associations between an exemplar and a social group are sufficient for evaluative generalization at the implicit level, but that propositional monitoring processes can reduce generalizations at the explicit level.

Drawing on earlier evidence for the differential effectiveness of affirmation versus negation foci in modulating automatic associations (e.g., Deutsch, Gawronski, & Strack, 2006), Gawronski, Deutsch, Mbirkou, Seibt, and Strack (2008) found that changes in implicit prejudice can be effected through repeated affirmation of positive associations, whereas repeated negation of negative associations increased implicit prejudice. These findings are in line with the APE model's assumption that negation of a proposition (e.g., "it is not true that old people are bad drivers") activates the association that underlies the proposition (i.e., the association between "old people" and "bad drivers"), thereby leading to ironic effects at the level of automatic associations (Wegner, 1994). In contrast, affirming

the opposite proposition (e.g., "old people are good drivers") activates an alternative association (i.e., the association between "old people" and "good drivers"), which results in intended responses at both the explicit and the implicit levels.

Contextualization

Although there is evidence that evaluations of minority members and social groups can be changed at the implicit and explicit level, evaluative responses may also vary as a function of the context. For instance, in a study by Wittenbrink, Judd, and Park (2001), spontaneous affective responses to Blacks were more favorable when they were presented in a positive context (e.g., family barbecue) than when they were presented in a negative context (e.g., graffiti wall). Based on findings like these, Wittenbrink et al. argue that evaluations of social groups can vary widely depending on the context in which they are presented, rather than being stable across contexts (see also Barden, Maddux, Petty, & Brewer, 2004). This assumption is in line with the APE model's contention that spontaneous affective responses can vary as a function of the pattern of associations that are activated, which in turn may depend on the particular context in which an object is encountered.

To understand the conditions that produce contextualization versus generalization of spontaneous affective reactions, Rydell and Gawronski (2009) conducted a series of studies in which they investigated the formation of context-dependent attitudes. Their results showed that newly formed implicit evaluations generalized across different contexts when information about the attitude object was homogeneous. However, when prior information about the attitude object was subsequently challenged by evaluatively incongruent information, implicit evaluations became context-sensitive, such that they reflected the contingency between the valence of prior information and the context in which this information was encoded. Importantly, when the available information about the attitude object was heterogeneous across different contexts, novel contexts still elicited implicit evaluations that reflected the valence of the initial experiences with the attitude object.

These results are important in the context of prejudice reduction. Specifically, Rydell and Gawronski's (2009) findings suggest that novel information about stigmatized groups may be "bound" to the particular context in which it is learned, such that this information is activated only in this particular context. As such, spontaneous affective responses in other contexts may still reflect the old, prejudiced evaluation (see also Gregg, Seibt, & Banaji, 2006). Thus, having positive experiences with stigmatized groups across multiple distinct contexts may be paramount in avoiding a specific contextualization of these experiences. If novel, positive experiences are limited to a particular context, it becomes more likely that the spontaneous affective reactions resulting from these experiences are limited to only this context, and that other contexts will still elicit the old, prejudiced response.

Subtyping

The mechanism of mentally creating "exceptions to the rule" by means of contextualization resembles the notion of subtyping, which refers to the phenomenon that counterstereotypical exemplars are often categorized in terms of subordinate groups, which leave the representation of the superordinate group unaffected (for a review, see Richards & Hewstone, 2001). Theoretically, exposure to multiple individuals who disconfirm a group stereotype should result in reduced application of the stereotype to the group. However, this expectation is often not met because exemplars that deviate from the stereotype tend to be subtyped. For example, meeting a wealthy and hardworking Black person may not result in the reduction of negative stereotypes about Blacks as a group (e.g., poor and lazy) because the subtype "Black professionals" can encapsulate these exceptions. Thus, individuals may continue to believe in the accuracy of the stereotype as applied to the group as a whole, despite their knowledge of several members who disconfirm this stereotype.

The propensity to subtype counterstereotypical exemplars is further exacerbated when these deviants are perceived to be highly atypical of the group as a whole (e.g., Hewstone, 1994; Kunda & Oleson, 1997; Weber & Crocker, 1983). Weber and Crocker (1983) examined the extent to which stereotypes about librarians and corporate lawyers could be changed through exposure to disconfirming exemplars. They found that stereotypes about the group tended to show stronger change when the disconfirming attributes were dispersed throughout the group, rather than being concentrated in a few members. Similarly, Hewstone (1994) determined that stereotypes about physics students tended to change only when the counterstereotypical exemplars were otherwise highly typical of the group. These results suggest that extreme deviants from the stereotype are less effective in producing stereotype change, as they are perceived to be highly atypical of the group and therefore are subtyped into subordinate categories.

One intriguing possibility from the perspective of the APE model is that exposure to counterstereotypical exemplars might have divergent effects on implicit and explicit evaluations of the group category. Preliminary evidence for this possibility comes from a study conducted by Dasgupta and Asgari (2004) which showed that exposure to famous female leaders led to a reduction in automatic stereotype activation on a gender-leadership IAT, but did not change stereotypic beliefs on explicit measures of female leadership capability. These findings suggest that the famous female leaders might have been subtyped at the explicit level, but that exposure to the counterstereotypical exemplars nevertheless produced a reduction in the associative strength of the stereotype. From the perspective of the APE

model, the associations underlying the stereotype may have been weakened following exposure to famous female leaders, but these associations may have been rejected on the basis that these exemplars are atypical of the group (for related evidence, see Moreno & Bodenhausen, 1999). These speculations are in line with the earlier reviewed findings by Ranganath and Nosek (2008) showing that generalizations from exemplars to groups occur immediately at the implicit level, whereas generalizations at the explicit level may be buffered by propositional monitoring processes.

Behavior-Based Interventions

So far, our discussion has focused primarily on prejudice reduction strategies that target people's mental representations of social groups. The basic assumption underlying this approach is that changes in mental representations lead to corresponding changes in overt behavior (see Gawronski & Sritharan, 2010). A potential alternative to such "mentalist" approaches are intervention strategies that start with people's behavior (see Olson & Stone, 2005). One example in this context is Festinger and Carlsmith's (1959) induced compliance paradigm, in which counterattitudinal behavior has been shown to reduce prejudice under certain conditions. However, even though induced compliance has been shown to be effective in reducing explicit prejudice (e.g., Gawronski & Strack, 2004; Leippe & Eisenstadt, 1994), it seems to be less effective in changing implicit prejudice (e.g., Gawronski & Strack, 2004; see also Wilson et al., 2000). In fact, from an evaluative conditioning perspective (see De Houwer et al., 2001), it would even seem possible that the negative feelings arising from cognitive dissonance become associated with the target group over time, such that the group is evaluated more negatively at the implicit level rather than positively. Other examples of behavior-based interventions are recent attempts to reduce prejudice by means of approach-avoidance training, in which participants are required to repeatedly respond with approach reactions to members of other groups (Kawakami, Phills, Steele, & Dovidio, 2007; see also Ito, Chiao, Devine, Loring, & Cacioppo, 2006). Research using approach-avoidance paradigms have consistently found training-related reductions in implicit prejudice. Explicit prejudice, in contrast, was largely unaffected by approach-avoidance training (e.g., Ito et al., 2006). From the perspective of the APE model (Gawronski & Bodenhausen, 2006a), these dissociations are consistent with the claim that dissonance-related reductions in prejudice are mediated by propositional processes, whereas approach-avoidance training influences prejudice via associative mechanisms (for a more detailed discussion, see Gawronski & Sritharan, 2010). Thus, even though the available evidence indicates that behavior-based interventions can be effective in reducing prejudice, their impact seems to be mediated by different underlying processes, thereby producing differential effects on explicit and implicit prejudice. As such, current behavioral intervention strategies still require further improvements to become an effective means of targeting both subtle and overt forms of prejudice simultaneously.

Future Directions

Although prior research shows that prejudice is malleable to some extent on both the implicit and the explicit level, some open questions remain. One particularly important task for applied research is the development of paradigms that simultaneously target multiple components of prejudice-related belief systems (see Figure 1). One promising approach in this objective might be diversity education. Rudman, Ashmore, and Gary (2001) found that participation in a diversity seminar on prejudice and intergroup conflict significantly reduced racial prejudice on both the explicit and implicit levels. In line with the implications of Gawronski et al.'s (2008a) integrative framework, changes in discrimination awareness and nonprejudicial goals were uniquely related to changes in explicit (but not implicit) prejudice, whereas changes in the liking of the course and the Black professor were uniquely related to changes in implicit (but not explicit) measures. The results of this study suggest that explicit prejudice reduction requires cognitive change through the stimulation of nonprejudicial goals and by increasing awareness of ongoing discrimination, whereas implicit prejudice reduction may require affective change through a decrease in fear of the outgroup and positive contact with outgroup members. Findings like these are very promising, although future research is needed to test whether the obtained reductions in prejudice remain stable over time.

An interesting question for basic research concerns the relationship between subtyping and contextualization. The reviewed findings by Rydell and Gawronski (2009) suggest that context-dependent learning operates at the level of associative processes, such that contextual cues may be incorporated in the associative representation of social groups during learning, which subsequently influence the affective responses that are elicited in a given context (see also Barden et al., 2004; Wittenbrink et al., 2001). This associative mechanism differs from the propositional mechanism that presumably underlies subtyping, in which counterstereotypical exemplars are excluded as "exceptions to the rule" at the propositional level, despite their inclusion in the associative representation of the group. This speculation is supported by the reviewed findings by Ranganath and Nosek (2008) and Dasgupta and Asgari (2004), who found that exemplar evaluations quickly generalized to the group at the implicit level, whereas generalization at the explicit level was buffered by propositional monitoring processes (see also Moreno & Bodenhausen, 1999). Future research comparing the processes underlying contextualization and subtyping effects may help to further

clarify the conditions under which the two mechanisms do and do not buffer prejudiced evaluations from counterattitudinal information.

Conclusion

Social psychological research has found that overt forms of prejudice have declined over the past few decades, but that subtle manifestations still remain prevalent in the general population. In the present article, we illustrate the usefulness of the APE model (Gawronski & Bodenhausen, 2006a, 2006b, 2007) in providing deeper insights into (1) the processes underlying overt and subtle forms of prejudice, and (2) the relative effectiveness of different strategies to reduce prejudice. The general conclusion that can be drawn from our discussion is that the effectiveness of prejudice reduction strategies depends on the particular components of prejudice-related belief systems that are targeted by these strategies. Because prejudice-related belief systems include multiple components that vary in their sensitivity to different types of influences, the most effective strategies are the ones that target all of the relevant components simultaneously. In this endeavor, the APE model may serve as a useful guide to identifying individual strategies that may be combined to achieve successful reductions of prejudice that remain stable across multiple contexts and over extended periods of time.

References

Amodio, D. M., & Devine, P. G. (2005). Changing prejudice: The effects of persuasion on implicit and explicit forms of race bias. In T. C. Brock, & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd ed., pp. 249–280). Thousand Oaks, CA: Sage.

Barden, J., Maddux, W. W., Petty, R. E., & Brewer, M. B. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology, 87,* 5–22.

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

Bonilla-Silva, E., & Forman, T. A. (2000). "I am not racist but . . ." mapping White college students' racial ideology in the USA. *Discourse & Society*, *11*, 50–85.

Brochu, P. M., Gawronski, B., & Esses, V. M. (2010). *The integrative prejudice framework and different forms of weight prejudice: An analysis and expansion*. Manuscript submitted for publication.

Brochu, P. M., Gawronski, B., & Esses, V. M. (2008). Cognitive consistency and the relation between implicit and explicit prejudice: Reconceptualizing old-fashioned, modern, and aversive prejudice. In T. G. Morrison, & M. A. Morrison (Eds.), *The psychology of modern prejudice* (pp. 27–50). Hauppauge, NY: Nova Science Publishers.

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, *87*, 546–563.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*, 642–658.

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

De Houwer, J., Thomas, S., & Baeyens, F. (2001) Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127,* 853–869.

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91,* 385–405.

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

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–52). New York: Elsevier.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23,* 316–326.

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring whites. *Journal of Experimental Social Psychology, 45,* 590–593.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25*, 603–637.

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

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.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row Peterson.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203–210.

Franco, F. M., & Maass, A. (1999). Intentional control over prejudice: When the choice of the measure matters. *European Journal of Social Psychology, 29,* 469–477.

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behavior? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19,* 285–338.

Fryer, R. G. (2007). Guess who's been coming to dinner? Trends in interracial marriage over the 20th century. *Journal of Economic Perspectives*, *21*, 71–90.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). San Diego: Academic Press.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes:

Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*, 23–30.

Gawronski, B., & Bodenhausen, G. V. (2006a). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731.

Gawronski, B., & Bodenhausen, G. V. (2006b). Associative and propositional processes in evaluation: Conceptual, empirical, and meta-theoretical issues. Reply to Albarracín, Hart, and McCulloch (2006), Kruglanski and Dechesne (2006), and Petty and Briñol (2006). *Psychological Bulletin, 132,* 745–750.

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25,* 687–717.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44,* 370–377.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15*, 485–499.

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44,* 1355–1361.

Gawronski, B., Peters, K. R., Brochu, P. M., & Strack, F. (2008). Understanding the relations between different forms of racial prejudice: A cognitive consistency perspective. *Personality and Social Psychology Bulletin, 34,* 648–665.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216–240). New York, NY: Guilford.

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40,* 535–542.

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

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

Hewstone, M. (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology, 5,* 69–109.

Hofmann, W., Gschwendner, T., Nosek, B. A., & Schmitt, M. (2005). What moderates explicit-implicit consistency? *European Review of Social Psychology, 16,* 335–390.

Ito, T. A., Chiao, K. W., Devine, P. G., Loring, T. S., & Cacioppo, J. T. (2006). The influence of facial feedback on race bias. *Psychological Science, 17,* 256–261.

Jacobson, C. K. (1985). Resistance to affirmative action: Self-interest or racism? *Journal of Conflict Resolution*, *29*, 306–329.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*, 137–146.

Kaiser, C. R., Drury, B. J., Spalding K. E., Cheryan, S., & O'Brien, L. T. (2009). The ironic consequences of Obama's election: Decreased support for social justice. *Journal of Experimental Social Psychology, 45,* 556–559.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology*, *55*, 893–905.

Kawakami, K., Phills, C., Steele, J., & Dovidio, J. (2007). (Close) distance makes the heart grow fonder: Improving implicit attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology, 92,* 957–971.

Kunda, Z., & Oleson, K. C. (1997). When exceptions prove the rule: How extremity of deviance determines the impact of deviant exemplars on stereotypes. *Journal of Personality and Social Psychology, 72,* 965–979.

Leippe, M. R., & Eisenstadt, D. (1994). Generalization of dissonance reduction: Decreasing prejudice through induced compliance. *Journal of Personality and Social Psychology*, *67*, 395–413.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego: Academic Press.

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution*, *25*, 563–579.

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

Moreno, K. N., & Bodenhausen, G. V. (1999). Resisting stereotype change: The role of motivation and attentional capacity in defending social beliefs. *Group Processes and Intergroup Relations, 2,* 5–16.

Olson, J. M., & Stone, J. (2005). The influence of behavior on attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 223–271). Mahwah, NJ: Erlbaum.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421–433.

Olson, M. A., & Fazio, R. H. (2009). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 19–63). New York: Psychology Press.

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

Petty, R. E., Cacioppo, J. T., Strathman, A. J., & Priester, J. R. (2005). To think or not to think: Exploring two routes to persuasion. In T. C. Brock, & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd ed., pp. 81–116). Thousand Oaks, CA: Sage.

Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately, explicit attitude generalization takes time. *Psychological Science, 19,* 249–254.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review, 5,* 52–73.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology*, *81*, 856–868.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, *17*, 437–465.

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23,* 1118–1152.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37,* 867–878.

Sears, D. O., & Henry, P. J. (2003). The origins of symbolic racism. *Journal of Personality and Social Psychology, 85,* 259–275.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99,* 3–21.

Sniderman, P. M., & Carmines, E. G. (1997). Reaching beyond race. *Political Science and Politics*, *30*, 466–471.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220–247.

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

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*, 961–977.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, *101*, 34–52.

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

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

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81,* 815–827.

Rajees Sritharan

Social Cognition Laboratory
Department of Psychology
The University of Western Ontario
Social Science Centre
London, Ontario N6A 5C2
Canada
E-mail rsrithar@uwo.ca