

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Hendricks Symposium--Department of Political
Science

Political Science, Department of

October 2006

Evolutionary Model of Racial Attitude Formation Socially Shared and Idiosyncratic Racial Attitudes

Thomas Craemer
University of Connecticut

Follow this and additional works at: <https://digitalcommons.unl.edu/politicalsciencehendricks>



Part of the [Political Science Commons](#)

Craemer, Thomas, "Evolutionary Model of Racial Attitude Formation Socially Shared and Idiosyncratic Racial Attitudes" (2006). *Hendricks Symposium--Department of Political Science*. 6.
<https://digitalcommons.unl.edu/politicalsciencehendricks/6>

This Article is brought to you for free and open access by the Political Science, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Hendricks Symposium--Department of Political Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

An Evolutionary Model of Racial Attitude Formation: Socially Shared and Idiosyncratic Racial Attitudes

**Thomas Craemer
University of Connecticut
Department of Public Policy
1800 Asylum Avenue, 4th Floor
West Hartford, CT 06117-2697**

**Telephone: (860) 570-9344
Fax: (860) 570-9114**

e-mail: thomas.craemer@uconn.edu

September 30, 2006

This paper was prepared for presentation at the Hendricks Conference on Biology, Evolution, and Political Behavior, October 13-14, 2006, in Lincoln, Nebraska.

Abstract. A growing body of research in political science has uncovered evidence of a “split personality” among Americans when it comes to racial attitudes, whereby people express different attitudes in public than they personally hold. A common assumption is that people adjust their personal attitudes to conform to dominant social norms. At present, however, there is *no* theoretical model that could account for the emergence of racial norms that are at odds with people’s personal attitudes. This paper proposes a simple neural model of racial attitude formation that makes an important distinction between *socially shared* and *idiosyncratic* racial attitudes. *Socially shared* attitudes reflect evaluations that are culturally transmitted and may not necessarily represent an individual’s personal views. In contrast, *idiosyncratic* attitudes represent a sense of interpersonal ‘chemistry’ that may be at odds with dominant social norms. A computational model based on Kimura’s (1983) Neutral Theory of Evolution predicts that *socially shared* racist attitudes may be able to coexist with, and eventually be replaced by, more favorable *idiosyncratic* racial attitudes.

In order to investigate racial attitudes unencumbered by considerations of social desirability and ‘political correctness,’ this study augments traditional survey measures with a number of reaction time based measures of non-conscious racial attitudes. *Socially shared*, non-conscious attitudes are measured using *implicit racial priming* based on a lexical decision task. *Idiosyncratic* non-conscious attitudes are measured using a timed trait rating procedure to measure feelings of *implicit closeness* towards African Americans and White Americans.

Experimental results (N=555) support the predictions derived from the computational model. They suggest that *socially shared* racial attitudes (as measured by implicit priming) are biased in a pro-White and anti-Black direction, even among non-White participants. This bias is interpreted as a subconscious remnant of old racist norms. On the *idiosyncratic* level, however, an entirely different picture emerges. Attitudes that are measured by the timed trait rating procedure are much more favorable toward African Americans. A multivariate analysis suggests that implicit closeness to Blacks drives support for race related policies such as affirmative action. Thus, *idiosyncratic* racial attitudes may be able to overcome the lingering effect of *socially shared* racist attitudes. The implications of the theoretical model and the empirical findings of this study are discussed and future research projects are proposed.

1. Introduction

A growing body of research in political science and social psychology has uncovered evidence of a “split personality” among Americans when it comes to racial attitudes (e.g., Devine 1989, Terkildsen 1993, Fazio et al. 1995, Greenwald et al. 1995, Kuklinski et al. 1997, Berinsky 2004, Feldman & Huddy 2005). People appear to voice different attitudes in public than privately when given the opportunity to express their personal views anonymously (e.g., Kuklinski et al. 1997).

This discrepancy is often interpreted as a social desirability effect among White¹ Americans who engage in self-monitoring. According to this interpretation, self-monitoring White respondents may adjust their old-fashioned, unfavorable views of African Americans to a new, pro-black norm of ‘political correctness’. In order to control for this social desirability effect, some researchers use Snyder and Gangestad’s (1986) self-monitoring scale (e.g., Terkildsen 1993, Berinsky 2004, Feldman & Huddy 2005). Other researchers attempt to measure unfavorable attitudes directly outside the respondents’ awareness using reaction time measures (Devine 1989, Fazio et al. 1995, Greenwald et al. 1995). They generally find a powerful pro-White and anti-Black bias among their White respondents on the non-conscious (implicit) level even among respondents who express favorable views on the conscious (explicit) level. Reviewing a large volume of evidence from the Implicit Association Test (IAT) in different domains of explicit and implicit attitudes, ranging from race and ethnicity to gender and age stereotypes, Greenwald et al. (2002) detect a general “empirical dissociation between the two types of measures” (Greenwald et al. 2002, p. 18). Social desirability explanations of this dissociation seem to imply that explicit attitudes are more susceptible to social norms than implicit ones. This may lead to the interpretation that implicit attitudes more faithfully represent an individual’s ‘personal’ attitudes. A radically different interpretation is implied by the model of dual attitudes proposed by Wilson et al. (2000). According to this model, implicit and explicit attitude-measures tap different aspects of an individual’s attitudes both of which may be influenced by personal feelings or social norms. The main difference is that implicit attitudes tend to reflect attitudes that have been rehearsed for a longer period of time and have become automatic. Such automatic responses require no conscious thought while newer attitudes require conscious effort. According to their

¹ In order to emphasize the socially constructed character of the race concept names of racial and ethnic groups are capitalized in this paper even if they refer to colors (e.g., Black, White, Black Americans, and White Americans).

interpretation, explicit attitudes may be just as ‘genuine’ as implicit ones, and they compare the rehearsal process to motor skills such as playing the piano or playing tennis. The conscious attempt to rehearse a new musical piece or a new serve cannot be interpreted as a disingenuous attempt at yielding to social norms, but as a genuine desire to play well. Similarly, favorable attitudes towards African Americans on the explicit level may represent a genuine desire at adopting a positive attitude, rather than a superficial attempt to satisfy social norms.

Whether social norms lead individuals to publicly misstate their personally held attitudes, or whether they infuse individuals with a genuine desire to adopt and rehearse new attitudes, neither interpretation offers a scenario that could explain how these social norms may emerge. This paper provides a simple theoretical model of racial norms evolution that is based on a few simple assumptions about neural organization and social communication. This model will be explained in the following section. A number of hypotheses are derived from the model and empirically tested based on a sample of N=555 college students in section 3. Finally, section 4 presents simulation results based on a computational version of the racial norms evolution model and compares the patterns observed in the simulations to the patterns observed in the student experiment.

2. A Simple Model of Racial Norms Evolution

The model of racial norms evolution proposed in this paper combines intra-personal properties of neural brain organization with inter-personal properties of social communication processes. On the intra-personal level, it assumes that sensory perceptions of internal body states and sensory perceptions of external stimuli are processed in different brain regions (see section 2.1). The model further assumes that both types of sensory perception, internal as well as external, are

subjected to a process of Hebbian learning (Donald O. Hebb 1949), whereby repeated rehearsal leads to automaticity. This process is equivalent to the rehearsal process described in Wilson's et al. (2000) dual attitude model and its neural basis will be described in greater detail in section 2.2. Finally, on the inter-personal level, the model of racial norms evolution is inspired by Motoo Kimura's (1983) theory of neutral evolution. This principle predicts the emergence of dominant norms by random drift even in the absence of selective advantages. This aspect of the model will be described in section 2.3.

2.1 Idiosyncratic and Socially Shared Attitudes

The *idiosyncratic vs. socially shared* distinction is based on the assumption that sensory perceptions of internal body states (internal stimuli) are processed in different brain regions than sensory perceptions of environmental events (external stimuli). While the latter can be observed by a number of individuals at the same time, the latter is perceived only by the individual him or herself. When people communicate it is easier to reach agreement about external stimuli than about internal ones. In addition, communications by others (whether in verbal form, in body language, or other symbols) enter the individual's brain as external stimuli and are assumed to be easier to communicate to others, since they are already stored in a communicable format. In contrast, internal body states enter the brain as diffuse sensations that are more difficult to express to others. Due to high levels of interconnectivity within the brain, these two regions are not assumed to be isolated, but rather to be interconnected in a peculiar form: Idiosyncratic perceptions of internal body states can be expressed in a communicable format with probability $p(\text{idiosyncratic})$. This probability depends on the level of difficulty translating internal body

sensations into a communicable format². It may be easier, for example, to express the internal body state of ‘feeling hungry’ than the complex sensation of feeling a ‘sense of chemistry’ with a complete stranger.

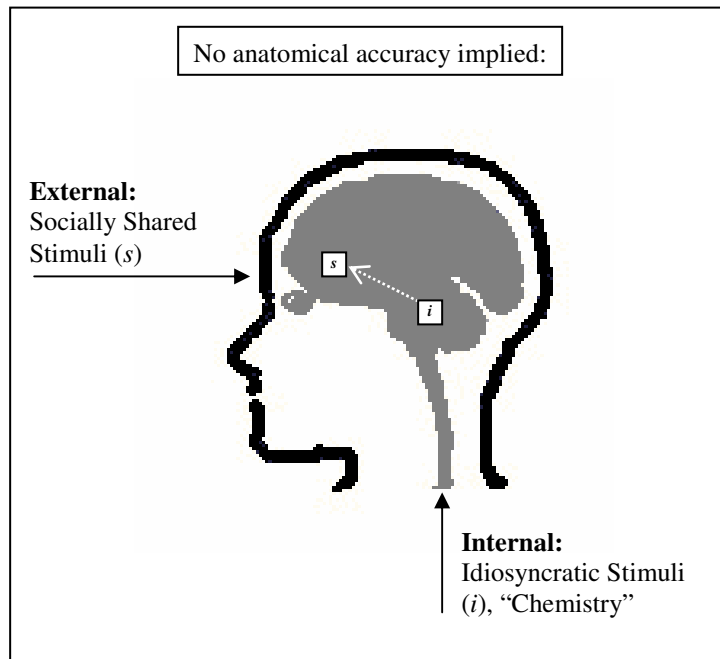


Figure 1: Distinction between Idiosyncratic (*i*) vs. Socially Shared Attitudes (*s*). The white arrow indicates the probability that an idiosyncratic attitude will be translated into an externally communicable format, $p(\text{idiosyncratic})$.

It is important to note that the expression of idiosyncratically perceived internal body states is a one way street. External perceptions of environmental stimuli, including communications from others, are assumed to be unable to influence internal body states. This is not an arbitrary assumption, but is rooted in the consideration that internal body states are likely to be the exclusive domain of an individual and cannot be directly determined by external stimuli. The same external stimulus may elicit very different body reactions in two different individuals. For

² Section 4 provides simulation results based on the model of racial norms evolution in which $p(\text{idiosyncratic})$ takes on various values.

example, a seafood connoisseur and an individual that is allergic to seafood may be able to talk about “seafood” based on a socially shared understanding of language, however, it will be associated with pleasant internal body states in the connoisseur, and with highly unpleasant ones in the allergic individual. No amount of emphasis on the part of the connoisseur of how delicious seafood is will be able to eliminate the unpleasant body state in the allergic individual. Figure 1 represents the spatial distinction between idiosyncratic attitudes (*i*) based on the perception of internal body states and socially shared attitudes (*s*) based on external stimulation. It is important to note that *no anatomical accuracy is implied*; the spatial *separation*, not the exact location, is of importance for the model proposed here.

2.2 Implicit and Explicit Attitudes

In line with Wilson’s et al. (2000) dual attitude model, the model of racial norms evolution assumes that frequently repeated (well rehearsed) thoughts, feelings, or motor functions are processed significantly faster than not so frequently repeated ones. This rehearsal effect is based on the principle of Hebbian learning (Hebb 1949). Donald O. Hebb (1949) formulated this principle based on the observation that the synaptic gap between two connected neurons tends to grow narrower with repeated simultaneous activation of the two neurons. The narrower the gap, the faster the signal transmission between the two neurons. Thus, frequently repeated thoughts, feelings, or motor functions are executed significantly faster than ‘new’ thoughts, feelings, or motor functions. The principle of Hebbian learning leads to associative learning in artificial neural nets and may provide a neural basis to Wilson’s et al. (2000) dual attitudes model as well as to the implicit-explicit attitude distinction made in the racial norms evolution model proposed here. The principle of Hebbian learning suggests that both types of racial attitudes discussed in

section 2.1 – *idiosyncratic* as well as *socially shared* – should become automatic after frequent activation (rehearsal). Thus, both types of racial attitudes should be detectable on the unconscious (implicit) level outside of an individual’s conscious control. This leads to a two-by-two classification scheme of racial attitudes distinguishing (1) *implicit idiosyncratic attitudes*, (2) *implicit socially shared attitudes*, (3) *explicit idiosyncratic attitudes*, and (4) *explicit socially shared attitudes*. Possible measurement methods for these four types of racial attitudes are discussed in section 3.

2.3 Random Norms Evolution

In addition to the two intra-personal processes described in sections 2.1 and 2.2, the model of racial norms evolution contains an inter-personal element that links various individual neural networks in a social network. Within this network, neighbors communicate with one another and exchange their socially shared attitudes (denoted s in figure 1). The fact that they exchange socially shared attitudes rather than idiosyncratic ones is based on the fact that these attitudes are already stored in a ‘communicable’ format while idiosyncratic attitudes require translation into such a format (see section 2.1). When two neighbors communicate, one neighbor is randomly designated as ‘persuader’ and the other as ‘persuadee.’ The persuadee is assumed to adopt the socially shared attitude of the persuader. Since each individual has exactly the same probability of being designated ‘persuader’ or ‘persuadee,’ no selection mechanism for a particular attitude is built into the model that could explain why a given attitude becomes socially dominant. Nonetheless, a process invariably occurs by which one out of any number of equal alternatives eventually emerges as a socially dominant majority attitude. This process, referred to in biological theory as ‘random drift,’ has been mathematically analyzed by Motoo Kimura (1983)

in his theory of neutral evolution.

Since this evolutionary model is new to the social sciences, a brief description of its biological origin is in place. It was developed by Kimura (1983) to explain evolutionary phenomena that cannot be explained by the Darwinian principle of natural selection. It became necessary when molecular genetics in the late 1960s, to their own surprise, encountered unequal distributions of selectively neutral synonymous alleles. Synonymous alleles are different DNA-sequences, that code for the same protein. Since an organism's selective advantage relies on the proteins it is composed of, these synonymous DNA-sequences are indistinguishable from one another by natural selection. This puzzle was solved by the strange dynamic of the Neutral Theory of Evolution that demonstrates how a purely random process will inevitably lead to distinct patterns in which one neutral alternative will dominate and eventually replace other equal alternatives. Kimura (1983) states: "The neutral theory asserts that the great majority of evolutionary changes at the molecular level ... are caused not by Darwinian selection but by random drift of selectively neutral ... mutagens" (Kimura 1983, S. xi). By virtue of the fact that Kimura's theory emphasizes the formative power of random mutational processes, it has become the dominant theory of evolution on the molecular level. Interestingly, the use of random evolution models in political science precedes Kimura's neutral theory. In 1968 William N. McPhee (1968) proposed a "Campaign Simulator" (p. 169) based on similar ideas and his colleagues Jack Ferguson and Robert B. Smith applied a related model to voting behavior. Due to the popularity of Darwinian models in social science applications, however, the power of Kimura's (1983) neutral evolution remains largely unexplored in contemporary political science. The main advantage of Kimura's (1983) random evolution model is its maximal parsimony which makes it attractive as a null model against which more restrictive theories could potentially be

tested.

When all three elements of the racial norms evolution model are taken together – the distinction between idiosyncratic vs. socially shared attitudes described in section 2.1, the distinction between implicit and explicit attitudes described in section 2.2 (see also Wilson et al. 2000), and the phenomenon of ‘random drift’ based on Kimura’s (1983) Theory of Neutral Evolution – a number of predictions can be derived. First, due to the dynamic of Kimura’s (1983) Neutral Theory Evolution, we should observe a single attitude to become a dominant social norm even in otherwise random communication processes. Based on the idea that socially shared and idiosyncratic attitudes may be processed in different brain regions (see section 2.1), the expectation naturally follows that over time idiosyncratic attitudes and socially shared ones should diverge for most individuals on topics that are widely discussed in the population. Further, attitudes (whether idiosyncratic or socially shared) that have been held by the individual for a longer period of time should tend to become automatic due to the process of Hebbian attitude rehearsal described in section 2.2 (compare also Wilson et al. 2000). Section 3 will discuss possible measures of implicit idiosyncratic and implicit socially shared attitudes, and it will present empirical findings based on a college student study (N=555). Section 4 will compare the patterns observed in the student study to simulation results based on a computational version of the Racial Norms Evolution model described in this section.

3. Empirical Measures and Empirical Results (N=555)

The difference between the four types of racial attitudes distinguished in section 2 may be illustrated by the historical example of Senator Strom Thurmond. His example suggests widely varying attitudes on the implicit and explicit idiosyncratic levels on one hand, and the implicit

and explicit socially shared ones on the other. In a Dallas radio address in September 1948, Thurmond voiced his opposition to integration, bluntly warning that integrationists “will find ... that in the lunchrooms, restrooms, recreation rooms, they will be compelled by law to mingle with *persons and races which all their lives they have, by free choice, avoided in social and business intercourse*” (cited in Stroud 2003, emphasis added TC). It came therefore as a surprise to many when it became known that Essie Mae Washington-Williams was the breathing result of social intercourse between her white segregationist father Thurmond and her African American mother Carrie Butler. On December 15th, 2003, Thurmond’s family announced: “As J. Strom Thurmond has passed away and cannot speak for himself, the Thurmond family acknowledges Ms. Essie Mae Washington-Williams’ claim to her heritage. We hope this acknowledgment will bring closure for Ms. Williams” (cited in Mattingly 2003).

Interestingly, on a personal level Thurmond appears to have acknowledged his African American daughter at the same time he publicly uninvited Virgin Island Governor Hastie with the following racist statement: “Governor Hastie knows that *neither he nor any other Negro will ever be a guest at the Governor’s house* in Columbia so long as I am Governor” (Thurmond on Oct. 25, 1948 cited in Stroud 2003, emphasis added TC). At the same time he secretly received his African American daughter Essie Mae Washington at the Governor’s Mansion (Stroud 2003). Further, he expressed his personal-social ambivalence in 1998 stating: “I’ve always held kindly to the black people ... But it was just the custom and the law that they were separate, and if I held otherwise, I would have been in violation of the law” (cited in Stroud 2003). That his first acknowledged fatherhood had been in violation of the very anti-miscegenation laws that Thurmond had so desperately fought to maintain flies in the face of any attempt to explain political behavior rationally. It is, however, consistent with a neural model that allows different –

even opposing – racial attitudes on the idiosyncratic and socially shared levels.

His initial attraction to Carrie Butler could be described as a result of idiosyncratic ‘chemistry’ resulting from the perception of internal body states and may well have been non-conscious (implicit idiosyncratic attitude). In addition, however, Thurmond seems to have consciously acknowledged his personal feelings (explicit idiosyncratic attitude) when he invited his daughter to the Governor’s Mansion. His use of the term “Negro” with a negative connotation suggests the simultaneous non-conscious presence of negative, socially shared, racial associations (implicit socially shared attitude) and his conscious and demonstrative un-invitation of Virgin Island Governor Hastie suggests conscious endorsement of these negative attitudes (explicit socially shared attitude).

For the purpose of empirical study, a number of measurement methods exist that may be able to tap into the different types of attitudes distinguished here. Aron et al. (1991) developed a trait-based reaction time task to measure non-conscious feelings of closeness among individual partners in close relationships (implicit idiosyncratic attitudes). This method has been successfully applied to measure implicit feelings of closeness toward social groups (Coats et al. 2000, Smith & Henry 1996) and will be described in greater detail in section 3.1. Explicit idiosyncratic attitudes are more difficult to measure in an empirical setting, especially if they deviate from social norms. Due to their private nature, people may be reluctant to express them to a stranger in a survey interview. In the future, it may be possible to apply in-depth interviewing by trusted friends to obtain measures of these feelings, but they are likely to remain the most elusive to empirical study. In order to measure non-conscious aspects of socially shared attitudes, racial priming measures are applied in this study (implicit socially shared attitudes). These measures rely on association of word meaning (positive or negative), and are likely to be

socially shared as an integral part of the communication process. The implicit racial priming measure applied for this study will be described in greater detail in section 3.2. Finally, explicit socially shared attitudes may be obtained by traditional survey measures. For the purpose of this study, questions like “How close do you feel towards African Americans?” and “How close do you feel towards White Americans?” were used, as were standard Feeling Thermometer ratings. Survey responses allow participants sufficient time to consciously monitor their responses, and in a politically sensitive area, such as race, it is likely that some respondents will adjust their responses to conform to dominant social norms. In the following two sections, the two main implicit measurement procedures that were obtained for this study will be explained: implicit closeness towards African Americans and White Americans (section 3.1), and Implicit Racial Priming (section 3.2).

3.1 Measuring Idiosyncratic Racial Attitudes Implicitly

Idiosyncratic measures of closeness towards African Americans and White Americans were obtained utilizing a timed trait self-rating procedure developed by Arthur Aron and his collaborators (1991). This measure operationalizes the idea that feelings of closeness may originate from an overlap of representations for the self and others in an individual’s mind. Their reaction time based non-conscious measure was originally developed to measure feelings of closeness between partners in individual relationships and has been successfully applied to the group-level by Eliot R. Smith and Susan Henry (1996), as well as Susan Coats and her collaborators (2000). In both studies, the timed trait rating procedure was used to measure subjective closeness between the self and a non-political ingroup (a sorority or fraternity). Coats et al. (2000) also investigated whether the implicit closeness measure would correlate with

explicit paper-and-pencil measures of closeness. They found a strong correspondence between conscious and non-conscious measures, a finding that does not seem surprising since the ingroups and outgroups they used (sororities and fraternities) appear innocuous from a social desirability point of view. Coats et al. (2000) write: “The advantages of implicit measures are obvious. They are not subject to self presentational or social desirability concerns. In addition, because implicit measures tap nonconscious, uncontrolled cognitive elements, they are less subject to demand characteristics” (Coats et al. 2000, p. 313). This makes the application of this non-conscious (implicit closeness) measure particularly attractive to the sensitive area of race and politics.

The timed trait self-rating task proceeds in two steps, an initial trait survey, and, after a distracter task, the actual timed self-rating procedure (see figure 2). The first step is required for classification purposes (for a detailed description see below) while the actual non-conscious closeness measure is obtained in the second step. In the following paragraphs, both steps of the procedure will be described in detail. Readers who are less interested in the technical details of the measure may skip ahead to section 3.3 for a summary of the results.

In the first step of the procedure (see figure 2), each individual participant is asked to rate each of 90 personality trait words³ as descriptive of the self, of “white Americans as a group,” and of “African Americans as a group.” Responses are given on a seven-point scale ranging from (1) “not at all” to (7) “extremely” descriptive with option (4) labeled as “neutral.” Based on these ratings, each trait is classified as either matching between the individual and a given group, or as mismatching. Note, although the meaning of the trait words themselves is likely to be socially shared, the question whether any given trait is descriptive of the respondent is highly idiosyncratic. Further, the self- and group ratings in this first step of the procedure are not used to

³ Thirty of the trait words have a negative connotation, thirty are neutral, and thirty positive, based on ratings provided by Anderson 1968, see Aron et al. 1991.

measure closeness, they serve strictly for classification purposes of ‘matching’ and ‘mismatching’ traits between the self and a group in the mind of the respondent. The actual measure of closeness occurs in the second step of the procedure in the timed self-rating task that will be described below.

Implicit Closeness Measure

Based on Aron et al. (1991)

- **Step (1) Trait Survey:** Rating Self, African Americans, and Whites on 90 Traits:
 "CONSIDERATE" Is this trait descriptive of (you as an individual / African Americans as a group / white Americans as a group)?
 1 = "Not at all"; 4="Moderately" 7 = "Extremely"
- **Step (2) Timed Trait Rating Task:** (after a distracter task and a break) Rating Self Only:
 Ask yourself this question:
 'Does this trait describe ME as an individual?'
 "CONSIDERATE"
 Yes, No.

Figure 2: Schematic representation of implicit closeness measurement procedure

In order to classify matching and mismatching traits, questionnaire responses are dichotomized with responses from (1) through (3) coded as ‘not descriptive’ and (5) through (7) as ‘descriptive’⁴. The top panel of table 1 gives an overview over the classification of matching (*M*) and mismatching (*M*) traits based on the responses to the initial trait survey. The number of matching traits (*A+B* in table 1) and mismatching ones (*C+D*) varies from one individual to the next. Thus, if an individual rates a given trait as descriptive of the self and a given group, a self-group match is recorded (*T_a* in the top panel of table 1). The same is true if the individual rates a given trait as non-descriptive of the self and non-descriptive of the group (*T_b* in the top panel of table 1).

⁴ Following Coats et al. (2000) neutral responses (4) are treated as missing information. No significant difference in missing information occurred between racial groups in the sample (see section 3).

Table 1: Computing Implicit Closeness Scores from Timed Trait-Self Rating Responses			
Step 1: Trait Survey – Classifying Matching (M) and Mismatching (M) Traits (T)			
Trait	Descriptive of self?	Descriptive of Group?	Pattern
Trait T_a of A Traits	Yes	Yes	Match M_x
Trait T_b of B Traits	No	No	Match M_y
Trait T_c of C Traits	Yes	No	Mismatch M_x
Trait T_d of D Traits	No	Yes	Mismatch M_y
Trait T_e of E Traits	Undecided	Undecided	Neutral (excluded)
Whereby $A + B + C + D + E = 90$ traits from Aron et al. (1991).			
Step 2: Timed Trait Rating Task – Reaction time (t) to correctly recognize trait (T)			
Reaction Time Score t	“Does this trait describe <i>me</i> as an individual?”	Mean Reaction Time	
$t_{M_x} = \sum_{a=1}^A t_a$	t_a : reaction time to recognize* T_a as self-descriptive	$\bar{t}_M = \frac{t_{M_x} + t_{M_y}}{A + B}$	
$t_{M_y} = \sum_{b=1}^B t_b$	t_b : reaction time to recognize* T_b as non-self descriptive		
$t_{M_x} = \sum_{c=1}^C t_c$	t_c : reaction time to recognize* T_c as self-descriptive	$\bar{t}_M = \frac{t_{M_x} + t_{M_y}}{C + D}$	
$t_{M_y} = \sum_{d=1}^D t_d$	t_d : reaction time to recognize* T_d as non-self descriptive		
*) only reaction times of correctly recognized traits are used to compute implicit closeness scores, totals $A, B, C,$ and D are adjusted accordingly.			
Step 3: Implicit Closeness Score towards either African Americans or White Americans			
<p>Implicit Closeness Score i_G towards group G :</p> $i_G = \bar{t}_M - \bar{t}_M$ <p>If the participant feels close to group G, the average reaction time for mismatching traits should be greater than that for matching traits ($\bar{t}_M > \bar{t}_M$) and the implicit closeness score i_G should be positive, else it should be close to zero or negative if the participant feels neutral or distant .</p>			

If the individual rates a given trait as descriptive of the self but not of the group or non-descriptive of the self but descriptive of the group a self-group mismatch is recorded (see

entries for traits T_c and T_d in the top panel of table 1). This procedure is applied to establish match and mismatch patterns for both groups under investigation, African Americans and White Americans. Thus, each trait is classified as a self-Black match, a self-Black mismatch, a self-White match, or a self-White mismatch. The center panel in table 1 shows how the distinction of matches (M) and mismatches (M) allows comparing average reaction times for matching and mismatching trait words in the subsequent timed self-rating task.

The timed self-rating task is the most crucial component of the non-conscious closeness measure. Each of the 90 trait words appears on the computer screen and the participant is asked to indicate as quickly as possible whether each word is self-descriptive or not. The instructions read: “Ask yourself this question: ‘Does this trait describe ME as an individual?’” followed by a trait word, e.g., “CONSIDERATE” (see figure 2). The participant then presses a button labeled as “Yes” or “No” as quickly as possible to record the response. What renders the timed trait rating measure non-conscious is the fact that the timed trait description only refers to the self. No reference is made to groups at this point. The psychological phenomenon that makes this procedure viable as a non-conscious measure of closeness is the curious fact that distinct facilitation and inhibition patterns occur for groups the individual feels close to while no such patterns occur for groups the individual does *not* feel close to. If, in the mind of a respondent, the self shares a trait with a close group, the trait is significantly faster identified as self-descriptive (facilitation). If the self differs from a close group on a trait, the trait is significantly slower identified as self-descriptive (inhibition). No such facilitation or inhibition effects occur for groups that the individual does not feel close to.

To compute implicit closeness scores toward each group (see center and bottom panels in table 1), reaction times (t) for all traits that are matching for the self and a given group are

averaged ($\overline{t_M}$) and subtracted from the average reaction times for all traits that are mismatching between the self and that group ($\overline{t_M}$). If the individual feels neutral towards the group (neither close nor distant) there should be no observable reaction time difference between matching and mismatching traits and the implicit closeness score should be close to zero. If the individual feels close to the group, matching traits should be recognized as self-descriptive significantly faster than mismatching traits and the implicit closeness score should be positive. The implicit closeness score should take on negative values if the individual feels distant to the group on the non-conscious level.⁵

The characteristic facilitation and inhibition effects that occur for close others and groups are interpreted by Aron et al. (1991) as indicating an overlap between the self-representation and the representation of others in the mind of an individual, “an actual overlap or confusion of cognitive structures” (Aron et al. 1991, p. 249). They write: “A possible explanation of the ... effect is that the cognitive structure of the self overlaps with the cognitive structure about the other ... Thus when a trait is descriptive of self but not other, there is a bit of confusion in deciding whether it actually represents the self” (Aron et al. 1991 p. 248).

3.2 Measuring Socially Shared Racial Attitudes Implicitly

In order to measure socially shared implicit racial attitudes, an implicit priming procedure was employed. The methodology of implicit priming was originally developed by James H. Neely (1977) and adapted for the purpose of measuring racial attitudes by Greenwald et al. (1995), as well as Fazio et al. (1995). The Implicit Association Test (IAT) developed by Greenwald et al. (1995) and Fazio's et al. (1995) racial priming method differ in a number of respects, but they

⁵ In order to reduce the inevitable skewness of reaction time measures, response times shorter than 300ms or longer than 2000ms are excluded from this computation (Coats et al. 2000, pp. 308-309).

share the idea that the positive or negative meaning of a prime word that flashes up on the participant’s computer screen is associated with the meaning of an otherwise unrelated target word. The participant is asked to indicate as quickly as possible whether the target word has a positive or negative meaning by pressing a button. Table 2 lists the prime and target words used for this study. In order to exclude conscious control on the part of the participant, prime words were displayed for a mere 20ms, too fast for conscious recognition. Thus, for participants, only the target words were consciously visible on the computer screen, appearing after a brief flash. The target words were selected from M.M. Bradley and P.J. Lang’s (1999) List of Affective Norms for English Words (ANEW) and are listed in table 2 along with their mean valence ratings⁶.

Table 2: Racial Priming: Prime and Target Words

Prime Words:	Target Words	Mean Valence	St. Dev.
“Black”	Love	8.72	0.70
	Joy	8.60	0.71
	Friendly	8.43	1.08
	Win	8.38	0.92
	Success	8.29	0.93
“African American”	Funeral	1.39	0.87
	Cancer	1.50	0.85
“White”	Rejected	1.50	1.09
	Sad	1.61	0.95
“White American”	Death	1.61	1.40

Target words and mean Valence ratings for target words from Bradley & Lang 1999

In order to reduce the inevitable skewness associated with raw reaction time measures, all responses shorter than 250ms and exceeding 2500ms were eliminated (truncation). The truncated raw reaction time measures were log-transformed to further reduce skewness. The transformed reaction time measures were then converted to facilitation scores by subtracting the reaction times

⁶ Five positive and five negative words were chosen from this list of 1036 normed words. Race un-related words were chosen with a frequency of at least F=25 and valence ratings greater than 8 for positive words and less than 2 for negative ones.

associated with racial prime-target pairs from neutral prime-target pairs. This was done for black and white prime words, as well as negative and positive target words separately. The neutral prime consisted of a white background with the character string “#####” displayed at the center of the computer screen. The facilitation score obtained in this way can be interpreted as the *relative acceleration* of responses following racial primes compared to neutral ones.

If t stands for the reaction time in milliseconds following a neutral prime and f stands for the facilitation score, then t / e^f gives the reaction time in milliseconds following a racial prime. For example, if a neutral prime takes about 600ms and the facilitation score is .0168, the reaction time for the racial prime can be obtained by computing $(600)/(e^{.0168}) = 590$. This represents a 10ms increase in reaction speed following the racial prime compared to the neutral one. The advantage of these relative facilitation scores over Fazio et al.’s (1995) raw facilitation scores is the fact that they eliminate individual differences in overall reaction time by looking at the relative value. (Non-truncated and non-log transformed raw facilitation scores show similar patterns at slightly lower levels of statistical significance).

3.3 Results Based on 555 College Students

A sample of N=555 undergraduate students participated in this study for credit at the Department of Political Science at Stony Brook University during the fall semester of 2003 and during the spring term of 2004. The demographic profile of the sample is not representative of the United States but it closely reflects the racial and ethnic composition of the undergraduate student body at Stony Brook. Fifteen percent of the sample consisted of international students (non-U.S. Citizens of any race) while 85 percent consisted of U.S. Citizens. Of these, 47 percent self-identified as White non-Hispanic, 10 percent as Black non-Hispanic, 11 percent as Hispanic

of any race, 24 percent Asian non-Hispanic, and 8 percent chose the residual ‘other’ category.

Each participant filled in a self-administered computer questionnaire in the Behavioral Labs of Stony Brook University’s Political Science Department. The computer questionnaire was programmed using Inquisit, a software package that allows precise reaction time measurements by controlling the computer’s task-prioritizing functions. This prevents programs running in the background from distorting reaction time measures. The study consisted of a political survey including questions about feelings towards social groups (closeness questions and feeling thermometer ratings) and a number of standard survey items on policy issues and demographic questions (including ideology, party ID, ethnicity and race). The study contained the two sets of implicit reaction time measures described in sections 3.1 and 3.2. A racial policy score was computed from four standard racial policy items (see table 7 in the appendix). This summary score was coded so that greater numbers represent more liberal views and smaller numbers more conservative views on issues such as affirmative action and government aid to African Americans.

To investigate the construct validity of the implicit and explicit racial attitude measures, they are entered as predictors into a maximum likelihood model of Pro-Black Policy Support in table 3. The variable ‘Motivation to Control Prejudice’ represents a control for social desirability consisting of a three item scale by Fazio et al. (1995). This scale contains questions that gauge how important it is to an individual to appear unprejudiced. The last column lists the impact of each independent variable, from the smallest observed value to the largest and allows for comparison among predictor variables. The explicit survey question “How close do you feel towards African Americans / White Americans” (Explicit Black / White Closeness in table 3) emerges as the most powerful predictor of racial policy preferences, whereby explicit closeness

towards African Americans predicts more liberal and explicit closeness towards whites more conservative opinions (both reaching or approaching significance at $p=.01$). Even after controlling for explicit closeness, implicit feelings closeness towards African Americans predict liberal opinions on race related policies at the $p<.01$ level of significance. While this idiosyncratic implicit measure exerts a powerful and significant effect of roughly the same magnitude as the standard 7-point ideology scale, the effect of the socially shared implicit racial priming measures are statistically indistinguishable from zero.

Table 3: Racial Policy Liberalism by Implicit and Explicit Attitudes

Racial Policy Liberalism ¹	Coef.	s.e.	p(z)	Impact*
Implicit Black Closeness	0.004	0.002	0.009	2.489
Implicit White Closeness	-0.002	0.002	0.220	-1.139
Implicit Black Priming	0.397	1.560	0.799	0.647
Implicit White Priming	-0.320	1.704	0.851	-0.401
Explicit Black Closeness	1.105	0.164	0.000	3.315
Explicit White Closeness	-0.441	0.180	0.014	-1.323
Conservative Ideology	-0.413	0.110	0.000	-2.478
Republican PID	-0.757	0.358	0.034	-0.757
Prejudice Monitoring	0.098	0.031	0.002	1.760
White non-Hispanic	-0.610	0.515	0.236	-0.610
Black non-Hispanic	0.339	0.646	0.599	0.339
Hispanic (any race)	0.668	0.607	0.271	0.668
Asian non-Hispanic	0.512	0.543	0.346	0.512
Constant	10.779	1.017	0.000	-
Log Likelihood	-608.613			
Wald χ^2 (13df)	209.870			
Prob > χ^2	0.000			
N	416			

¹) Racial Policy Liberalism: 4-item summary scale, wording see table 7 in the appendix. Estimation Method: Maximum Likelihood; Significant coefficients shaded to facilitate interpretation. *) Impact: impact of each independent variable going from the smallest observed value to the largest one.

Republican partisanship is predictive of racially conservative opinions ($p<.05$) and the ‘Motivation to Control Prejudice’ is predictive of racially liberal opinions ($p<.01$). The latter coefficient suggests that participants who state that it is important to them to appear unprejudiced

may over-exaggerate their support for race targeted policies such as affirmative action. Interestingly, once implicit and explicit feelings of closeness toward racial groups are controlled for, racial and ethnic group membership ceases to exert any significant influence. The fact that implicit idiosyncratic feelings of closeness towards African Americans appear to be large and significant predictors of racial policy liberalism supports the construct validity of this implicit idiosyncratic measure. The fact, however, that racial priming measures appear to be unrelated to racial policy liberalism at first blush casts doubt on their validity as a measures of implicit socially shared racial attitudes. At closer inspection, however, their lack of predictive power and their lack of correlation with explicit measures of socially shared racial attitudes may not be so surprising. This lack of correlation is consistent with a large body of literature on the Implicit Association test, and it follows from the prediction of the Racial Norms Evolution Model presented in section 2. Thoroughly rehearsed socially shared attitudes should be universally shared due to the process of ‘random drift’ described by Kimura (1983, see section 2.3). As a universally shared attitude, it should take on a near constant value resulting in zero-correlations.

This lack of correlation between implicit priming measures has been described in the literature on the Implicit Association Test (IAT, Greenwald et al. 1995). Greenwald et al. (2002) review a large volume of IAT evidence and find that dissociations between implicit and explicit attitudes using this measure have been confirmed in the area of race and ethnicity and have also been identified in the area of gender and age (Greenwald et al. 2002, p. 18). While social desirability concerns may provide a plausible explanation for dissociation in this socially sensitive domain, the same explanation does not seem plausible as an explanation of low correlations in the area of flowers, musical instruments, insects, and weapons that formed the initial basis for the IAT measure in Greenwald et al. (1995). The lack of correlation in the latter

domain would seem to be much more consistent with an interpretation of implicit word-association measures (IAT and Racial Priming) as tapping socially shared aspects of word meaning. Further support can be gleaned from implicit attitudes of the racial IAT from non-White participants. A number of studies suggest that non-White participants, especially African Americans, display as much pro-White and anti-Black bias in their implicit word associations as their White American counterparts. Using the standard computerized version of the IAT, for example, Dasgupta et al. (2000) find pro-White and anti-Black reaction time patterns despite the fact that more than half (53 percent) of their sample is composed of non-White respondents. Jost et al. (2004) reviews findings from the internet version of the IAT and finds that “members of low-status minority groups (including African Americans) commonly fail to exhibit ingroup bias and show preferences for higher-status outgroups – even when these preferences are soundly rejected at an explicit, conscious level” (Jost et al., 2004, p. 893). Nosek et al. (2002) describe this contrast: “strong explicit liking reported by Black respondents for their own group stands in sharp contrast to performance on the implicit measure” (Nosek et al. 2002, p. 105). Ashburn-Nardo et al. (2003) obtained IAT measures of N=80 African Americans and found “the IAT effect was significantly different from zero and in a negative direction ... underscoring the degree to which many black participants in our sample exhibited relatively negative ingroup associations” (Ashburn-Nardo et al. 2003, p.73) and conclude: “Indeed, mounting evidence suggests that many ... black individuals do hold outgroup-favoring associations at the implicit level” (Ashburn-Nardo et al. 2003, p. 80).

In order to compare the results of this study to the studies cited above, all racial attitude variables obtained in this study were converted into scores ranging from a pro-White extreme to a pro-Black one with zero representing the neutral midpoint (where attitudes towards African

Americans and Whites are equal). To make the measures comparable, they were converted to z-scores and midpoint was preserved by adding the z-score for the neutral zero point to each original z-score. This z-score with neutral midpoint is denoted as $z' = z + (0 - \text{mean}) / \text{st.dev.}$ Figure 3 displays the z' -scores for the three racial attitude measures under consideration in this study. Bars flagged with two stars represent z' -scores that are significantly different from the neutral zero-point, whereby positive numbers represent comparatively more pro-Black attitudes and negative numbers comparatively more pro-White ones. The results suggest that implicit idiosyncratic attitudes (implicit closeness represented by black bars), despite their powerful predictiveness of racial policy preferences, do not significantly deviate from zero in any racial group under consideration. This is consistent with the idea of an idiosyncratic measure that should display variance between individuals, not between groups.

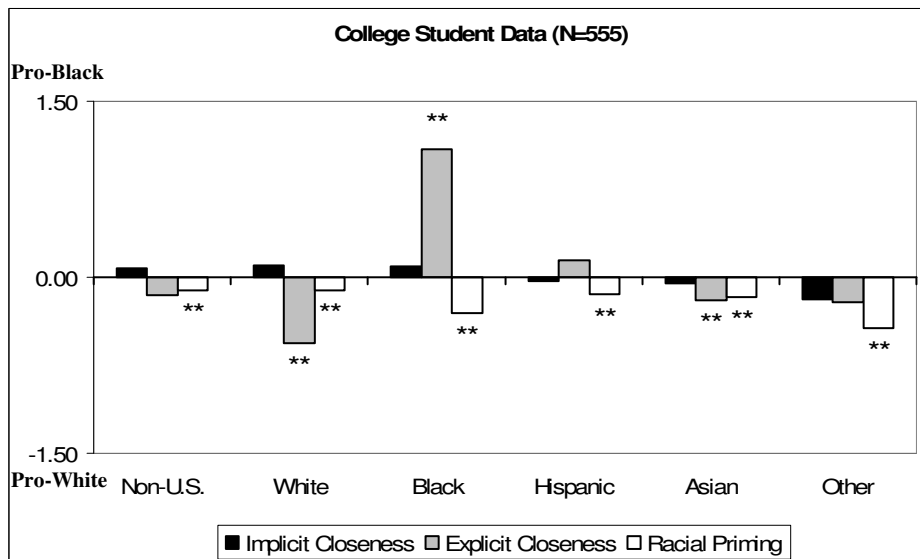


Figure 3: College Student Data by Racial Groups

In contrast, explicit socially shared attitudes measured by explicit closeness questions (grey bars) display systematic and large differences between groups. White and Asian American participants express significantly more pro-White than pro-Black attitudes, while African Americans express

overwhelmingly pro-Black attitudes (all deviations from the neutral zero-point are significant at $p < .01$). When the results of implicit socially shared measures are considered (Racial Priming represented by white bars), no group differences are visible, instead all groups display a highly significant ($p < .01$) pro-White and anti-Black bias. This is consistent with the findings from the IAT literature cited above and supports the that reaction time procedures based on word-associations may tap universally shared, not necessarily individually endorsed, aspects of racial attitudes. In their inaugural presentation of the IAT methodology, Greenwald et al. (1995) support this interpretation by demonstrating that “IAT measures were highly sensitive to evaluative discriminations that are well established in the connotative meaning structure of the English language” (Greenwald et al. 1995, p. 1469). The contrast that emerges, especially for African American participants, between their explicit socially shared attitudes (grey bar in figure 3) and their automatic word-associations (white bar) can be illustrated with a historical example: In his Address at the Conclusion of the Selma to Montgomery March on March 25th, 1965 Martin Luther King counter-intentionally invoked ingrained linguistic and racial associations while explicitly criticizing such stereotypes: “How long will prejudice blind the visions of men, *darken* their understanding, and drive *bright-eyed* wisdom from her sacred throne?” (King in Carson & Shepard 2001, p. 130, emphasis added TC). The explicit attitudes expressed give voice to the views of the Civil Rights Movement, while the figures of speech invoked implicitly represent ingrained elements of the overall linguistic culture.

The significance tests in table 4 give the differences in group means for the measures in figure 3 (ANOVA F -Test for Racial Group Differences) and the overall deviation from the neutral zero-point within the entire sample (t -Test for Deviation of Grand Mean from Scale Midpoint). As would be expected for an implicit idiosyncratic measure (Implicit Closeness), no

group differences emerge (see left hand panel in table 4) and no systematic racial bias can be detected (see right hand panel of table 4). In contrast, for the explicit measure of socially shared attitudes (Explicit Closeness), significant group differences emerge (left hand panel), and the sample mean is significantly and systematically biased in a pro-White and anti-Black direction (see negative t -test value in the right hand panel). Finally, for implicit socially shared attitudes (Racial Priming), no group differences can be distinguished (left hand panel), but a significant and systematic pro-White and anti-Black bias can be observed (see negative t -test value in the right hand panel of table 4). This is consistent with the universal IAT-effect observed in the IAT-literature (Greenwald et al. 1995, Greenwald et al. 2002, Dasgupta et al. 2000, Jost et al. 2004, Nosek et al. 2002, Ashburn-Nardo et al. 2003). In order to investigate the theoretical consistency of these empirical findings with the Model of Racial Norms Evolution presented in section 2, a computational version of the model is presented in the following section and simulated results are compared to the patterns observed in figure 3 and table 4.

Table 4: Group Norms (F -Tests) and Universal Norms (t -Tests) in Student Data (N=555)

	ANOVA F -Test for Racial Group Differences		t -Test for Deviation of Grand Mean from Scale Midpoint	
	ANOVA F	$p(F)$	t	$p(t)$
Implicit Closeness	0.876	0.497	0.821	0.413
Explicit Closeness	29.450	0.000	-4.468	0.000
Racial Priming	.990	0.423	-3.915	0.000

4. A Computational Model of Racial Norms Evolution

The computational model of Racial Norms Evolution simulates a population of the same size as the sample of college students in section 3 (N=555). Individuals are arranged on a two-dimensional grid of 15 x 37 cells. The racial composition of the simulated population reflects that of the college student sample in section 3 and group members are clustered to mimic

racially segregated living arrangements frequently encountered in the contemporary United States. Each individual is assumed to have two racial attitudes, one idiosyncratic, and one socially shared. Attitudes are represented by random numbers between zero and one, whereby zero represents extremely pro-White and one extremely pro-Black attitudes. The midpoint of the scale represents neutrality where attitudes towards Whites and Blacks are equal. At the beginning of the simulation process attitudes are randomly initialized from a uniform distribution between zero and one, and idiosyncratic as well as socially shared attitudes are identical. Throughout the simulation process, idiosyncratic attitudes are kept constant, while socially shared attitudes are subject to change through random persuasion. In each simulation round t , one individual is selected at random to be the ‘persuader’ I_t and one of the direct neighbors in the grid is randomly selected as the ‘persuadee’ J_t . The persuadee takes on the socially shared attitude of the persuader, so that $J_t = I_t$. There is a probability of taking on one’s own idiosyncratic attitude as socially shared attitude instead of the persuader’s attitude and this probability is denoted by $p(\text{idiosyncratic})$. This probability is set at different values during five simulations ranging from $p(\text{idiosyncratic})=0.00$, 0.25, 0.50, 0.75, to $p(\text{idiosyncratic})=1.00$. At one extreme, at $p(\text{idiosyncratic})=0.00$, individuals never express their idiosyncratic racial attitudes and are maximally susceptible to persuasion by others. At the other extreme where $p(\text{idiosyncratic})=1.00$ they express *only* their idiosyncratic attitudes and are entirely immune to persuasion by others. The implicit-explicit dimension is represented by time (the number of persuasion rounds). Applying the Hebbian interpretation of attitude rehearsal, attitudes that have been rehearsed for longer periods of time should become automatic (implicit). Thus, simulated attitudes after 50,000 persuasion rounds might represent explicit socially shared attitudes, while simulated attitudes after 300,000 rounds might represent implicit socially shared norms.

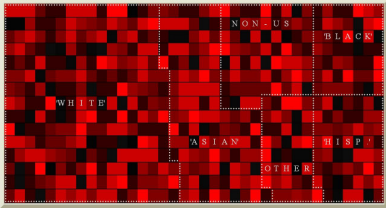
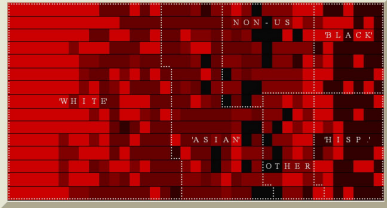
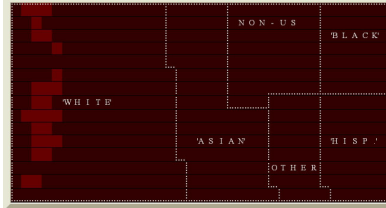
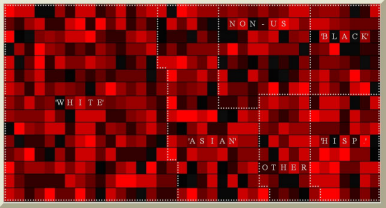
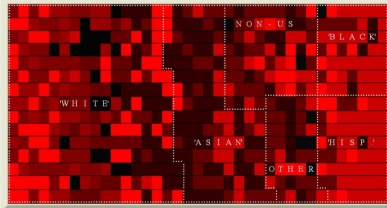
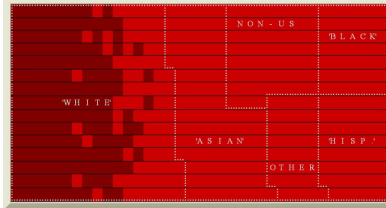
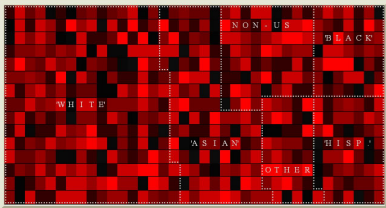
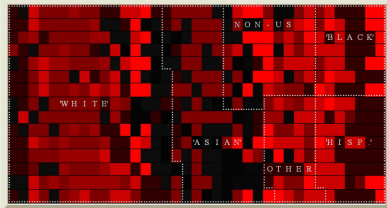
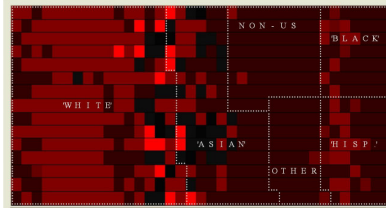
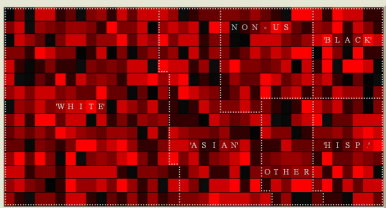
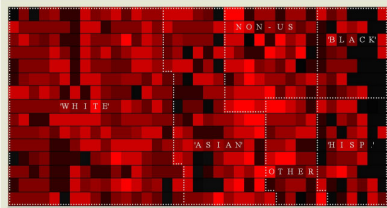
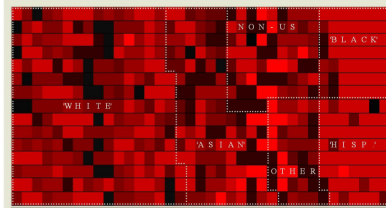
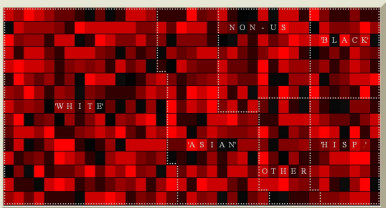
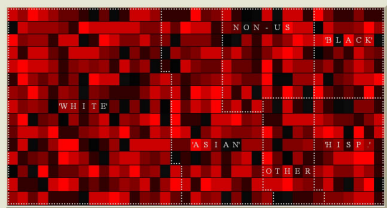
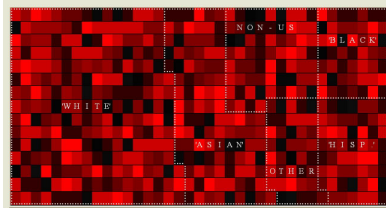
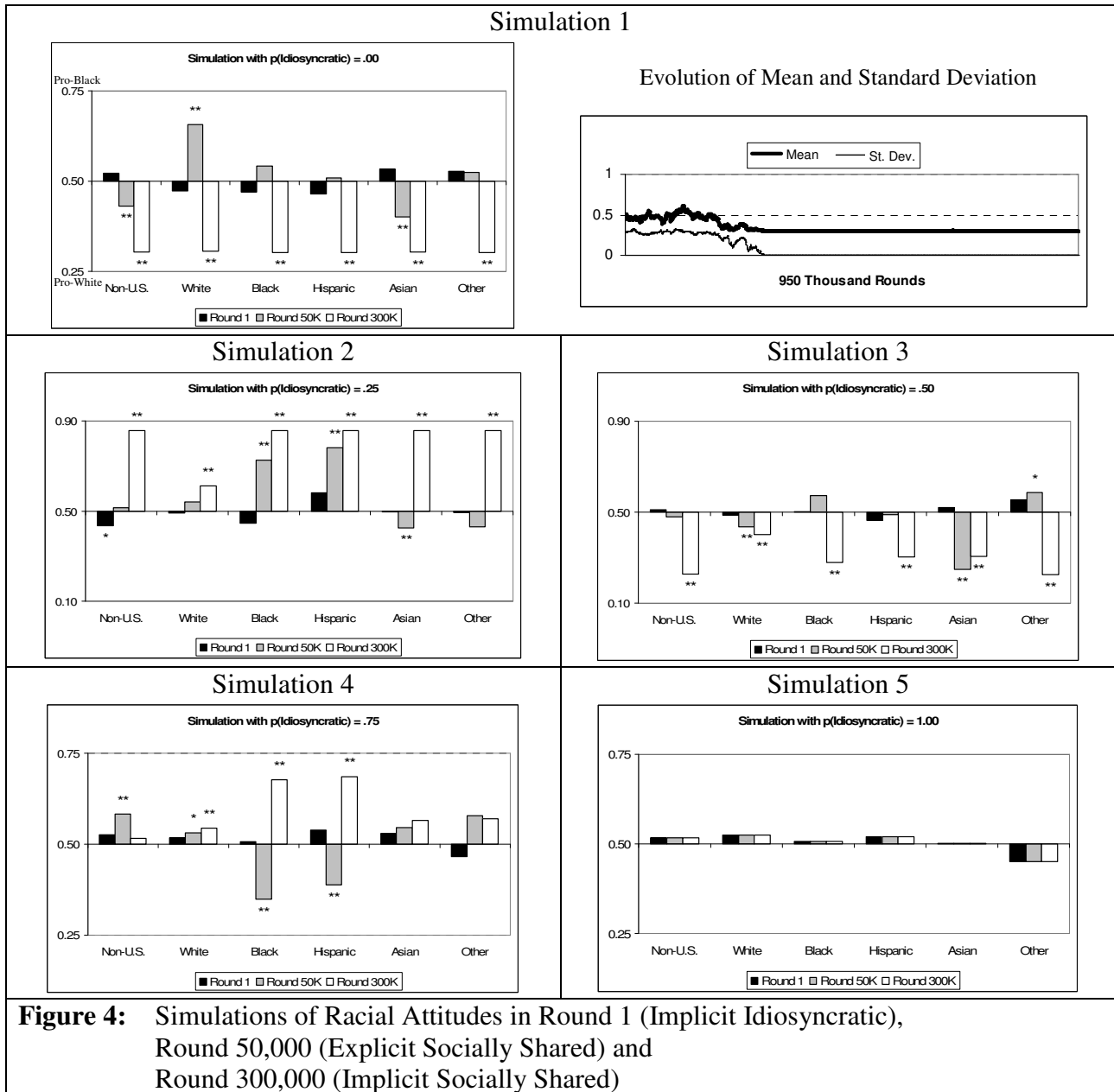
Table 5: Five Simulations of Racial Norms Evolution with Different Settings of $p(\text{idiosyncratic})^*$		
Round 1:	Round 50,000	Round 300,000
Simulation 1: $p(\text{idiosyncratic}) = 0.00$		
		
Simulation 2: $p(\text{idiosyncratic}) = 0.25$		
		
Simulation 3: $p(\text{idiosyncratic}) = 0.50$		
		
Simulation 4: $p(\text{idiosyncratic}) = 0.75$		
		
Simulation 5: $p(\text{idiosyncratic}) = 1.00$		
		
<p>*) $p(\text{idiosyncratic})$ represents the probability of an individual's idiosyncratic attitude being expressed. Each square in this 15 x 37 matrix represents a single individual ($N=555$); the number of individuals in clusters represent the number of individuals of the indicated racial or ethnic category in the college student experiment described in section 3. Attitudes are color coded from most pro-African American (red in color or light gray in gray scale representation) to most pro-White American (black in both color and gray scale).</p>		

Table 5 presents the results of the five simulations described above. Each square in the matrices represents the socially shared racial attitude, one individual color coded to represent the most

pro-African American attitudes as bright red, and the most pro-White American attitudes as black. For each simulation, the initial distribution at round 1 is displayed on the left, the results after 50,000 persuasion rounds in the middle, and after 300,000 persuasion rounds on the right. Idiosyncratic attitudes are omitted from table 5 since they are equal to the initial distributions on the left hand side and remain constant throughout the simulation process. This does not imply that idiosyncratic attitudes are viewed as immutable in reality, they are simply held constant for purposes of simulation in order to investigate the dynamics of persuasion on the socially shared level. In all simulations in table 5 in which $p(\text{idiosyncratic})$ is less than 1.00, clustering of socially shared attitudes is evident. That is, as long as social communication is permitted to have any influence at all, norms of varying strength emerge. This process is most pronounced in simulations 1 and 2 where the influence of idiosyncratic attitudes is relatively low (simulations 1 and 2 with $p(\text{idiosyncratic})=.00$ and $p(\text{idiosyncratic})=.25$ respectively). While a relatively pro-White and anti-Black norm emerges as dominant in simulation 1 (dark red or grey coloring), a relatively pro-Black and anti-White one emerges in simulation 2 (bright red or grey coloring). This process of random norms evolution emerges according to the dynamics of Kimura's (1983) Neutral Theory of Evolution and is only absent in simulation 5 in which the social influences are precluded by virtue of setting $p(\text{idiosyncratic})$ to 1.00. In simulations 3 and 4, with $p(\text{idiosyncratic}) \geq .50$, the norms formation is less pronounced and statistical analysis is required to detect the degree to which social norms emerge.

Figure 4 presents bar-graphs similar to the bar graphs for the college student sample in figure 3. For each of the racial groups, the bars represent mean deviations from the neutral midpoint of the simulated racial attitude scale, deviations in a positive direction representing more pro-Black attitudes (brighter red or gray in table 5) and deviations in a negative direction

representing more pro-White attitudes (darker red or gray in table 5). The top panel in figure 4 represents simulation 1 for which the $p(\text{idiosyncratic})$ was set at zero. Invariably, under this setting a single attitude survives and the simulation comes to an end once no further changes are possible. The bar chart shows initial (idiosyncratic) attitudes (black bars) not differing significantly from the neutral midpoint of the racial attitude scale. After 50,000 persuasion rounds (see grey bars) significant group differences have emerged, with some groups taking on norms that deviate from the neutral midpoint in a positive direction, and others in a negative direction. After 300,000 persuasion rounds, all groups have adopted a universally pro-White and anti-Black norm. The graph on the right hand side of the top panel in figure 4 plots the mean (bold black line) and the standard deviation (thin black line) of racial attitudes for the entire duration of simulation 1. It illustrates how the mean starts off at the neutral midpoint initially, veers off the neutral midpoint for a while, until it reaches an equilibrium state far off the neutral midpoint. As the norm crystallizes, the standard deviation gradually decreases over the first third of the evolution (about 300,000 rounds), and remains just above zero for the remainder of the simulation until the last individual adopts the social norm in round 950,000 and the standard deviation goes to zero. This is the process of random drift described in Kimura's (1983) Neutral Theory of Evolution. The only prediction that can be made with certainty in this random model is the fact that a norm will evolve – the direction of the norm is entirely unpredictable, as is the question how far away from the neutral midpoint it will fall. The process becomes indefinite once idiosyncratic attitudes supply a constant supply of new minority attitudes. A large number of simulations were run apart from the simulations displayed in table 5 to ensure that the results are not simply peculiar patterns, but predictably reoccurring patterns. When the bar graphs for the four simulations are compared for which $p(\text{idiosyncratic}) < 1.00$, the results are surprisingly similar.



In each case, initial idiosyncratic attitudes (black bars) are, on average, close to the neutral midpoint without significant group differences, after 50,000 persuasion rounds (grey bars), socially shared attitudes display significant group differences and bars point in different directions. Finally, after 300,000 persuasion rounds, white bars generally point in the same direction (universal norm) and are significantly different from the neutral midpoint for most

groups. The only exception is simulation 5 for which $p(\text{idiosyncratic})=1.00$ and social influences are ruled out by definition. In this case the initial idiosyncratic distribution remains constant, no group differences occur, and no universal norm emerges that is significantly different from the neutral midpoint.

Table 6: Group Norms (*F*-Tests) and Universal Norms (*t*-Tests) in Simulation Data

	ANOVA <i>F</i> -Test for Racial Group Differences		<i>t</i> -Test for Deviation of Grand Mean from Scale Midpoint	
	ANOVA <i>F</i>	<i>p</i> (<i>F</i>)	<i>t</i>	<i>p</i> (<i>t</i>)
Simulation 1				
Round 1	1.124	0.346	-0.393	0.695
Round 50,000	21.081	0.000	3.780	0.000
Round 300,000	29.773	0.000	-1174.014	0.000
Simulation 2				
Round 1	1.998	0.077	-0.734	0.464
Round 50,000	17.761	0.000	3.688	0.000
Round 300,000	153.641	0.000	38.176	0.000
Simulation 3				
Round 1	0.697	0.626	0.056	0.956
Round 50,000	15.565	0.000	-5.421	0.000
Round 300,000	19.867	0.000	-23.514	0.000
Simulation 4				
Round 1	0.398	0.850	1.440	0.151
Round 50,000	9.508	0.000	1.231	0.219
Round 300,000	7.232	0.000	7.808	0.000
Simulation 5				
Round 1	0.496	0.779	1.001	0.317
Round 50,000	0.496	0.779	1.001	0.317
Round 300,000	0.496	0.779	1.001	0.317

Table 6 presents statistics that corroborate this interpretation. Looking at the ANOVA-*F* tests on the left hand side, the results for simulations 1 through 4 show no significant group differences in initial round 1. After 50,000 rounds, significant group differences emerge and remain throughout round 300,000. The right hand side of table provides *t*-tests for the emergence of a universal norm. The patterns for simulations 1 through 4 mirror those of the student data in table 4 in many important respects – no significant universal norm exists in round 1 (idiosyncratic attitudes), while significant deviations from the neutral midpoint emerge after social communication has

taken place (explicit and implicit socially shared attitudes). The only simulation in which no norm evolves is simulation 5, for which social communication has been excluded by definition ($p(\text{idiosyncratic})=1.00$). This simulation experiment suggests that the random Norms Evolution process is a robust phenomenon that emerges as long as any persuasion is possible at all. Although the simulated process has been a highly simplified and parsimonious version of a real world process, the complex patterns that emerge display some surprising similarities to observed data.

5. Conclusions

The theoretical model of Racial Norms Evolution presented here combines two assumptions of neural organization with one assumption about social communication processes. The first of these assumptions holds that perceptions of internal body states ('chemistry' or idiosyncratic attitudes), and perceptions of external stimuli (socially shared attitudes) are processed in different (yet interconnected) areas of the brain. The second assumption holds that attitudes rehearsed for a longer period of time become automatic by a process of Hebbian learning (Hebb 1949, see also Wilson et al. 2000). This leads to a two-by-two classification of attitudes, (1) implicit idiosyncratic, (2) explicit idiosyncratic, (3) implicit socially shared, and (4) explicit socially shared, three of which were considered in this study (1, 3, and 4). The third assumption of the model holds that social communication can produce social norms by virtue of a process that has been described by Kimura (1983) as 'random drift'. The tendency for computational models of social communication processes to produce unanimity as a function of random rather than systematic processes has been noted with puzzlement by some social scientists. Andrzej Nowak et al. (1990), for example, criticize: "the implicit null hypothesis seemingly held by most social

psychologists is that group processes, if allowed to work themselves through to their conclusion, would lead to a final distribution of opinion ... with zero variance” (Nowak et al. 1990, p. 363). Due to the fact that public opinion research generally focuses on divisive issues rather than unanimous ones, this general tendency of random norms evolution has been dismissed by political scientists as an anomaly and several computational modelers have sought to limit this tendency by including restrictive elements in their models. Nowak et al. (1990), for example, propose a computer model in which stable local ‘pockets’ of dissent remain due to the fact that some members of the society have zero persuasiveness. Similarly, Robert Axelrod (1997) proposes a model for the dissemination of culture in which communication is a function of ‘similarity’ between two individuals and stable pockets of dissent remain if similarity between some individuals is set to zero: “If they are completely different, they will not even interact” (Axelrod 1997, p. 211). In practice, however, it is hard to imagine what ‘zero similarity’ between two people could mean, since they will at least share their common humanity. Similarly, it is hard to imagine what ‘zero persuasiveness’ means as long as a human being is able to communicate at all. Despite these rather strong and restrictive assumptions these models do not produce stable opinion splits close to the 50 percent mark typical for the ones identified in many public opinion surveys, but produce rather small pockets of dissenting views and sometimes none at all. This paper takes a different approach and allows for the possibility that this general tendency of random norms evolution may have a parallel in real world communication processes. It may help us explain the otherwise puzzling observation that people adjust their views to powerful social norms (e.g., Devine 1989, Terkildsen 1993, Fazio et al. 1995, Greenwald et al. 1995, Kuklinski et al. 1997, Berinsky 2004, Feldman & Huddy 2005). If we assume that social norms are simply an additive function of individual attitudes, no systematic differences should occur between average

idiosyncratic and average socially shared attitudes.

According to the college student experiment described in section 3, no systematic racial bias is observable at the level of implicit idiosyncratic measures (Implicit Closeness), while a significant anti-Black bias is observable at the level of both explicit and implicit socially shared attitudes (Explicit Closeness, and Racial Priming respectively). The simulation results presented in section 4 faithfully replicate these patterns as long as social communication is possible at all, i.e., as long as $p(\text{idiosyncratic}) < 1.00$. The model is maximally parsimonious, holding idiosyncratic attitudes constant, and allowing changes in socially shared attitudes due to random persuasion among neighbors. More sophisticated versions of the computational model could allow for random change in idiosyncratic attitudes, and it could allow for individual differences in $p(\text{idiosyncratic})$. While these changes might lead to more realistic results, it is noteworthy that a maximally parsimonious baseline model suffices to produce considerable similarity between observed and simulated racial attitude data. Due to its parsimony, the Model of Racial Norms Evolution presented here may serve as a null model against which more complex models can be tested.

Appendix:

Table 7: Components of the Racial Policy (RP) Dependent Variable	
RP 1:	Because of past discrimination, minorities should be given special consideration when decisions are made about hiring applicants for jobs - do you strongly agree, somewhat agree, somewhat disagree, or strongly disagree?
RP 2:	The government in Washington should make every possible effort to improve the socioeconomic position of Blacks and minority groups - do you strongly agree, somewhat agree, somewhat disagree, or strongly disagree?
RP 3:	Where would you place the government in Washington's efforts to improve the social and economic position of Blacks and other minority groups on a scale ... where 1= the government should not make any special effort, and 7 = the government should make every possible effort?
RP 4:	The government should not make any special effort to help Blacks and other minorities because they should help themselves - do you strongly agree, somewhat agree, somewhat disagree, or strongly disagree?
The responses are summed up and recoded so that greater numbers represent greater racial policy liberalism and smaller numbers greater racial policy conservatism. Source for question wording: NES, NBES, see Tate (1993)	

Literature:

- Anderson, N. H. (1968): Likableness Ratings of 555 Personality-Trait Words. *Journal of Personality and Social Psychology* 9: 272-279.
- Aron, Arthur, Elaine N. Aron, Michael Tudor, and Greg Nelson (1991): Close Relationships as Including Other in the Self. *Journal of Personality and Social Psychology* 60 (2): 241-253.
- Ashburn-Nardo, Leslie, Megan L. Knowles, and Margo J. Monteith (2003): Black Americans' Implicit Racial Associations and Their Implications for Intergroup Judgment. *Social Cognition* 21(1): 61-87.
- Axelrod, Robert (1997): The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, V. 41, N. 2, pp. 203-226.
- Berinsky, Adam J. (2004): Can We Talk? Self-Presentation and the Survey Response. *Political Psychology* 25 (4): 643-659.
- Bradley, M.M. and P.J. Lang's (1999): *Affective Norms for English Words (ANEW)*. Gainesville, FL. The NIMH Center for the Study of Emotion and Attention, University of Florida.
- Carson, Clayborne, and Kris Shepard, eds. (2001): *A Call to Conscience. The Landmark Speeches of Dr. Martin Luther King, Jr.* New York, NY: Warner Books.
- Coats, Susan, Eliot R. Smith, Heather M. Claypool, and Michele J. Banner (2000): Overlapping Mental Representations of Self and In-Group: Reaction Time Evidence and Its Relationship with Explicit Measures of Group Identification. *Journal of Experimental Social Psychology* 36: 304-315.
- Darwin, Charles, 1859, 1993: *The Origin of Species*. New York: Random House, Inc..
- Dasgupta, Nilanjana, Debbie E. McGee, Anthony G. Greenwald, and Mahzarin R. Banaji (2000): Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *Journal of Experimental Social Psychology* 36: 316-328.
- Devine, Patricia G. (1989): Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology* 56: 5-18.
- Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams (1995): Variability in Automatic Activation as Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline? *Journal of Personality and Social Psychology*, 69, 6: 1013-1027.
- Feldman, Stanely, and Leonie Huddy (2005): Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice? *American Journal of Political Science* 49 (1): 168-183.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwarz (1995): Measuring Individual Differences in Implicit Cognition: the Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464-1480.
- Greenwald, Anthony G., Mahzarin R. Banaji, Laurie A. Rudman, Shelly D. Farnham, Brian A. Nosek, and Deborah S. Mellott (2002): A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept. *Psychological Review* 109(1): 3-25.
- Hebb, Donald O. (1949): *The Organization of Behavior. A Neuropsychological Theory*. New York, NY: John Wiley.
- Jost, John T., Mahzarin R. Banaji, and Brian A. Nosek (2004): A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo. *Political Psychology* 25 (6): 881-919.

- Kimura, Motoo, (1983): *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Kuklinski, James H., Michael D. Cobb, and Martin Gilens (1997): Racial Attitudes and the 'New South'. *Journal of Politics* 59 (2): 323-349.
- Maison, Dominika, Anthony G. Greenwald, Ralph H. Bruin (2004): Predictive Validity of the Implicit Association Test in Studies of Brands, Consumer Attitudes, and Behavior. *Journal of consumer Psychology* 14(4): 4005-415.
- Mattingly, David (2003): Strom Thurmond's Family Confirms Paternity Claim. CNN Washington Bureau, 12/16/2003. Source: <http://www.cnn.com/>
- McPhee, William N. (1968): *Formal Theories of Mass Behavior*. New York, NY: The Free Press of Glencoe, a division of The MacMillan Company.
- Neely, James H. (1977): Semantic Priming and Retrieval from Lexical Memory. *Journal of Experimental Psychology: General* 1977, V106 N3, pp. 226-254.
- Nosek, Brian A., Mahzarin R. Banaji, and Anthony Greenwald (2002): harvesting Implicit Group Attitudes and Beliefs From a Demonstration Web Site. *Group Dynamics: Theory Research, and Practice* 6(1): 101-115.
- Nowak, A., J. Szamrej, and B. Latané (1990): From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact. *Psychological Review* Vol. 97, No. 3: 362-376.
- Smith, Eliot R., and Susan Henry (1996): An In-Group Becomes Part of the Self: Response Time Evidence. *Personality and Social Psychology Bulletin* 22 (6): 635-642.
- Snyder, Mark, and Steven W. Gangestad (1986): On the Nature of Self-Monitoring. *Journal for Personality and Social Psychology* 51 (1): 125-139.
- Stroud, Joseph S. (2003): Thurmond's Past Invites New Scrutiny. *The State*, South Carolina, 12/21/2003. Source: <http://www.thestate.com/mld/thestate/>
- Tate, Katherine (1993). *From Protest to Politics: The New Black Voters in American Elections*. New York: Russell Sage.
- Terkildsen, Nayda (1993): When White Voters Evaluate Black Candidates: The Processing Implications of Candidate Skin Color, Prejudice, and Self-Monitoring. *American Journal of Political Science* 37 (4): 1032-1053.
- Wilson, T. D., S. Lindsey, and T. Y. Schooler (2000): A Model of Dual Attitudes. *Psychological Review* 107: 101-126.