Easier Done Than Undone: Asymmetry in the Malleability of Implicit Preferences

Aiden P. Gregg
University of Southampton

Beate Seibt
Universität Würzberg

Mahzarin R. Banaji
Harvard University

Dual-process models imply that automatic attitudes should be less flexible than their self-reported counterparts; the relevant empirical record, however, is mixed. To advance the debate, the authors conducted 4 experiments investigating how readily automatic preferences for one imagined social group over another could be induced or reversed. Experiments 1 and 2 revealed that automatic preferences, like self-reported ones, could be readily induced by both abstract supposition and concrete learning. In contrast, Experiments 3 and 4 revealed that newly formed automatic preferences, unlike self-reported ones, could not be readily reversed by either abstract supposition or concrete learning. Thus, the relative inflexibility of implicit attitudes appears to entail, not immunity to sophisticated cognition, nor resistance to swift formation, but insensitivity to modification once formed.

Keywords: implicit, IAT, attitude, malleability

Social psychologists have long noted that self-reported measures of attitude, though useful and convenient tools, are vulnerable to several validity-impairing biases, such as self-presentation (Schlenker, 1975; Baumeister, 1982), self-deception (Greenwald, 1988; Paulhus, 1993), and self-ignorance (Converse, 1970; Nisbett & Wilson, 1977; Wilson, Dunn, Kraft, & Lisle, 1989). To get around these biases, social psychologists have long sought alternative, subtle means of attitude assessment (see Crosby, Bromley, & Saxe, 1980, for an early review). Their search has culminated most recently in the development of *implicit measures*, specialized techniques that capitalize on respondents' nondeclarative responses to attitude objects to illuminate respondents' automatic associations towards those objects.

Implicit measures are methodologically diverse (Bassili, 2001; Brauer, Wasel, & Niedenthal, 2000; De Houwer & Eelen, 1998; Greenwald & Banaji, 1995; Hetts, Sakuma, & Pelham, 1999; Koole, Dijksterhuis, & van Knippenberg, 2001; Vanman, Paul, Ito, & Miller, 1997; Von Hippel, Sekaquaptewa, & Vargas, 1997) but most commonly take the form of compatibility tasks in which targets, distracters, and responses vary in their semantic or evaluative congruency (Bargh, Chaiken, Govender, & Pratto, 1992; De Houwer, 2003; Fazio, Sanbomatsu, Powell, & Kardes, 1986; Glaser & Banaji, 1999; Greenwald, Draine, & Abrams, 1996; Greenwald, Klinger, & Liu, 1989; Greenwald, McGhee, & Schwartz, 1998; although also see Koole & Pelham, 2003). Such tasks are designed to make controlled responding difficult (Neely, 1977), and their purpose is often disguised to minimize respondent reactivity (Fazio, Jackson, Dunton, & Williams, 1995).

Defusing initial suspicions to the contrary (e.g., Bosson, Swann, & Pennebaker, 2000), implicit measures of attitude [1] have proven capable of predicting a range of important phenomena (Egloff &

Aiden P. Gregg, School of Psychology, University of Southampton, England; Beate Seibt, Department of Psychology, Universität Würzberg, Germany; Mahzarin R. Banaji, Department of Psychology, Harvard University.

Research reported in this paper was partially supported by grants from the National Institute of Mental Health (MH-57672), the National Science Foundation (SBR-9709924), and the Third Millennium Foundation and by a fellowship from Rockefeller Foundation's Bellagio Study and Conference Center to Mahzarin R. Banaji. Some of these data were reported at the 1999 annual meeting of the Midwestern Psychological Association in Chicago, IL.

We thank Johanna Zumer, Joe Bushill, and Anne-Marie Brister for their assistance in data collection. We thank Irene Blair, Yarrow Dunham, Jeff Ebert, Tony Greenwald, Brian Nosek, and Claire Hart for their helpful comments on a previous draft.

Correspondence concerning this article should be addressed to Aiden P. Gregg, School of Psychology, University of Southampton, Highfield Campus, Southampton SO17 1BJ, England, United Kingdom. E-mail: aiden@soton.ac.uk

[1] Although focus of our research is on automatic evaluative associations ("attitudes"), many of the points we made are likely to apply equally to automatic semantic associations ("beliefs") (e.g., Banaji & Hardin, 1996). We use the catch-all term "automatic attitude" to imply that our theorizing potentially straddles both types of association.

Journal of Personality and Social Psychology, 2006, Vol. 90, No. 1, 1–20
Copyright 2006 by the American Psychological Association 0022-3514/06/$12.00 DOI: 10.1037/0022-3514.90.1.1

Schmuckle, 2002; Greenwald, Nosek, & Banaji, 2003; Jellison, McConnell, & Gabriel, 2004; Koole & Pelham, 2003; Maison, Greenwald, & Bruin, 2001; Marsh, Johnson, & Scott-Sheldon, 2001; Nosek, Banaji, & Greenwald, 2002b; Teachman & Woody, 2003), most notably, spontaneous behavior that explicit measures fail to predict (Asendorpf, Banse, & Mucke, 2002; Perugini, 2004; Dovidio, Kawakami, & Gaertner, 2002; Dovidio, Kawakami, C. Johnson, B. Johnson, & Howard, 1997; Fazio et al., 1995; McConnell & Leibold, 2001; Neumann, Hülsenbeck, & Seibt, 2004; Spalding & Hardin, 1999). A recent meta-analysis (Poehlman, Uhlmann, Greenwald, & Banaji, 2004) revealed that the prognostic power of the Implicit Association Test (IAT; Greenwald et al., 1998), a leading implicit measure, was not far behind that of explicit measures overall and surpassed it in the reactive domain of stereotyping and prejudice. Such findings augur well for the construct validity of implicit measures as indices of automatic attitudes.

However, knowledge of what implicit measures predict can only provide partial insight into the nature of automatic attitudes. Full insight awaits the empirical elucidation of their antecedents, that is to say, of factors that lead automatic attitudes to emerge in the first place and cause them to change thereafter. A two-pronged enterprise, exploring antecedents in tandem with consequences, is highly characteristic of science in general, as well as of research on attitudes in particular (cf. Petty & Krosnick, 1995). In what follows, we make an empirical foray designed to redress the current paucity of research on the antecedents of automatic attitudes.

Antecedents of Automatic Attitudes

According to the classic dissociation model (Devine, 1989), automatic attitudes toward social groups form inexorably over time. By belonging to a culture, people cannot help being exposed to information, in the media and elsewhere, that links different social groups to positive and negative attributes. Because these links are repeatedly and chronically activated, their activation eventually becomes automatic for all members of that culture, requiring only that the social groups in question, or some representation of them, be perceived. If people later come to disagree, on principled grounds, with how a given social group is portrayed in their culture, they must inhibit the automatic associations that they have passively acquired during socialization and activate in their place the enlightened insights they have arrived at upon mature reflection. What distinguishes egalitarians from bigots, on this view, is not their automatic social attitudes, which are collectively shared and deeply ingrained, but rather their conscious social attitudes, which are individually entertained and readily malleable.

Subsequent empirical research has called into question some of the more provocative claims of the original dissociation model (Blair, 2002). Automatic attitudes are not invariably activated in everyone after all: factors such as attention (Castelli, Zogmaister, Smith, & Arcuri, 2004; Gilbert & Hixon, 1991; Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997) and motivation (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Spencer, Fein, Wolfe, Fong, & Dunn, 1998; Lepore & Brown, 1997; Sinclair & Kunda, 1999; Wittenbrink, Judd, & Park, 1997) play moderating and mediating roles. Nevertheless, one key postulate of the dissociation model remains viable, namely, that automatic attitudes generally show a greater resistance to change than their self-reported counterparts. In particular, the dissociation model proposes that people become egalitarian in their professed ideals before they become egalitarian in their underlying sympathies, a proposition consistent with much cross-sectional data on racial attitudes that have been assessed both explicitly and implicitly (Dovidio & Gaertner, 1998; Gaertner & Dovidio, 1986; Fazio, 2001).

Moreover, the dissociation model is joined by several cognate dual-process theories from which similar predictions about the relative malleability of automatic and self-reported attitudes could be derived (Chaiken & Trope, 2000; Sloman, 2002). Smith and DeCoster (1999) have postulated the existence of two complementary representational systems: a rule-based one, in which sudden transformations of serial representations (or symbols) occur, and an associative one, in which gradual transformations of connectionist representations (or weights) occur. Epstein and Pacini (1999) contend, as part of their cognitive–experiential self-theory, that the mind contains both a rational system, characterized by relative flexibility, and an experiential system, characterized by relative inertia. Finally, Wilson, Lindsey, and Schooler (2000) have drawn a distinction between attitudes that are consciously constructed online and attitudes that take the form of more enduring dispositions. Thus, there is ample theoretical precedent to postulate that automatic attitudes might be more stable than their self-reported counterparts. What is more, there is a good deal of supportive empirical evidence.

Empirical Evidence for the Stability of Automatic Attitudes

To begin with, automatic attitudes, unlike their self-reported counterparts, resist attempts at deliberate manipulation. Kim (2003) found that White participants directly instructed not to show automatic preferences while performing an Implicit Association Test (IAT; Greenwald et al., 1998) persisted in showing them. Moreover, naïve respondents tend not to spontaneously discover effective "faking" strategies (Banse, Seise, & Zerbes, 2001; Egloff & Schmuckle, 2002; Foroni & Mayr, in press; although also see Lowery, Hardin, & Sinclair, 2001) even if they could in principle deploy them with the benefit of instruction or hindsight (Blair & Banaji, 1996; Steffens, 2004). There is also direct evidence that automatic attitudes, as originally hypothesized, reflect the "introspectively unidentified. . .traces of past experience" (Greenwald & Banaji, 1995, pp. 8). In an elegantly designed study, Petty and Jarvis (1998) began by classically conditioning a preference in participants for one photographed face over another. Then, to induce a further preference based on perceived observer–target similarity, they led participants to believe that they shared more opinions in common with one of the people photographed (e.g., Eddie) than with the other (e.g., Phil). Several different groups of participants were created. In some, both attitude manipulations were designed to induce the same preference (e.g., both manipulations portrayed Eddie as preferable to Phil); in others, both manipulations were designed to produce contrary preferences (e.g., one manipulation portrayed Eddie as preferable to Phil, and the other, Phil as preferable to Eddie). At the end of the experiment, explicit ratings of Eddie and Phil were found to have been influenced only by perceived similarity. However, performance on an evaluative priming task (cf. Fazio et al., 1995) was found to have been additionally influenced by the face preferences classically conditioned at the outset of the study. This shows that, despite superficial appearances and explicit manipulations, earlier learning can subtly persist. Similar findings have been obtained using introspection as a manipulation, and rapid responding as an implicit measure (Wilson, Lindsey, & Schooler, 2000).

Further evidence for the robustness of automatic attitudes comes from recent work on cognitive dissonance (Festinger, 1957; Harmon-Jones & Mills, 1999). Standard counterattitudinal manipulations that reliably cause changes in self-reported attitudes have been found to leave implicit attitudes unaffected, consistent with the latter being more stable than the former (Gawronski & Strack, 2003; although also see McDell, Banaji, & Cooper, 2004). In addition, field data broadly consistent with such laboratory findings have also emerged. For example, Hetts et al. (1999) documented a stepwise rise in self-esteem among different generations of East Asian immigrants to the United States. The rise was manifested on explicit measures (self-report questionnaires) before being manifested on implicit measures (evaluative priming tasks). Finally, data from vast Web surveys show the attitudes toward age (*young* vs. *old*) and academic disciplines (*math* vs. *arts*) are comparable across the life span (Nosek, Banaji, & Greenwald, 2002a). This strongly suggests that at least some automatic attitudes are generationally stable.

Empirical Evidence for the Malleability of Automatic Attitudes

The picture is far from one-sided however. Much empirical evidence also attests to the remarkable malleability of automatic attitudes (for a review, see Blair, 2002). For example, Dasgupta and Greenwald (2001) found that White participants exposed to favorable exemplars of Black Americans and unfavorable exemplars of White Americans manifested weaker automatic preferences for their own race than did control participants; their overtly expressed racial preferences, however, did not change. The effect, sustained over several days and replicated in a different domain, suggests that automatic attitudes may be amenable to change even when their conscious counterparts are not. In addition, Lowery et al. (2001) found levels of automatic race prejudice can be moderated even by casual social encounters, although the effect depends on racial backgrounds of those involved and respondents” relative status (Richeson & Ambady, 2003; see also Banaji, 2002).

Other investigators (Foroni & Mayr, in press) have found that normative automatic preferences for flowers over insects can be significantly curtailed simply by having participants read a story presenting a fictional rationale for entertaining contrary preferences (i.e., flowers become radioactive and insects nutritious on a postapocalyptic earth). Blair, Ma, and Lenton (2001) established that automatic associations could be reduced even without exposure to any external sources of information. Participants who summoned up imagery at odds with gender stereotypes showed attenuated automatic biases relative to participants who summoned up neutral, stereotypical, or no mental imagery. This result held for several different implicit measures, thereby casting doubt on alternative explanations invoking response suppression or shifting response criteria.

Several additional studies have documented substantial shifts in automatic attitudes based on recent experiences, variations in context, or variations in physiological state (Gawronski, Walther, & Blank, 2004; Karpinski & Hilton, 2001; Lane, Mitchell, & Banaji, 2004; Mitchell, Nosek, & Banaji, 2003; Rothermund & Wentura, 2001; Seibt, Hafner, & Neumann, 2004). Finally, investigations that have teased apart internal consistency from temporal stability have found that underlying automatic attitudes show marked variability from one measurement occasion to another (Gregg & Sedikides, 2004; Steffens & Buchner, 2003).

The Present Research

What are we to make of this paradoxical picture? Some research suggests that automatic attitudes are relatively inflexible (either absolutely or relative to self-reported attitudes). However, other research, no less compelling, suggests that they are relatively malleable. Given these mixed results, it may be useful to adopt for the moment a pragmatic *perspectivist* approach (Banaji, 2002; McGuire, 1973). That is, it may be useful to regard the claim that automatic attitudes are inflexible and the claim that they are malleable as being true in some contexts, but false in others—each time in potentially illuminating ways. The ultimate resolution of the paradox must await the development of an integrative theory, one that specifies the chief boundary conditions under which automatic and self-reported attitudes exhibit one characteristic or the other (cf. Petty, 1997). However, given that such an integrative theory is probably a distant prospect, there are good grounds for adopting in the interim a research orientation designed to generate a broad body of findings that can representatively inform any subsequent integrative theorizing. That is to say, a bottom-up inductive approach, furnishing the raw materials for future theoretical edifices, may prove as useful as a top-down deductive approach, checking the solidity of theoretical edifices so far assembled. As Sherlock Holmes [2] once remarked to his faithful companion, Dr. Watson, “It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts” (Doyle, 1981, p. 165).

Bearing this in mind, we embarked on a coordinated series of laboratory studies—theory-testing in intent, but also reflecting an openness to alternative formulations implied by the data—to investigate the ease with which automatic and self-reported preferences for one imagined social group over another might be induced or undone. We derived our provisional theoretical starting-point from the classic dual-process models reviewed above, which are consensually regarded as implying that automatic and self-reported attitudes are affected to different degrees by cognitive processes of greater or lesser sophistication. In our view, dual-process models can also be more narrowly interpreted as implying that, whereas both self-reported and automatic attitudes may be responsive, in some measure, to what we term *abstract supposition* and *concrete*

[2] Not that Holmes shunned deductive reasoning; quite the reverse, indeed, given his line of work (although Conan Doyle most often portrays him as making inspired *inductions*). As Holmes remarked to Watson on another occasion: “How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?” (Doyle, 1981, p. 111). Good science, like good detective work, relies on a judicious combination of induction and deduction.

learning, self-reported attitudes should be relatively more responsive to abstract supposition, and automatic attitudes relatively more responsive to concrete learning.

Let us elaborate. We define concrete learning as the act of cognitively assimilating multiple pieces of information about the characteristics of an object or, alternatively, of assimilating the same piece of information multiple times. Thus, reading a detailed descriptive account of some object or undergoing a session of intensive associative conditioning (De Houwer, Thomas, & Baeyens, 2001) would both qualify as instances of concrete learning. Abstract supposition, in contrast, we define as the act of hypothetically assuming that an object possesses particular characteristics. Thus, entertaining the idea, out of the blue, that a novel object is X or ~X or that an existing object known to be X is in fact ~X (or vice versa), both qualify as instances of abstract supposition. The critical difference between the two is that, in the case of concrete learning, the characteristics of an object are implied by, or inferred from, an elaborate set or protracted series of prior instances, whereas in the case of abstract supposition, no such set or series of instances is available. In other words, the act of abstractly supposing that some state of affairs is the case involves entertaining cognitions that are *purely* formal and symbolic. Consequently, abstract supposition should be particularly well suited to activating explicit representations—namely, those that are "rule-based," "rational," and "constructed"—whereas concrete learning should be particularly well suited to activating implicit representations—namely, those that are "association-based," "experiential," and "dispositional." Now, if classic dual-process models are correct, then explicit representations ought to register most clearly on self-reported measures of attitude and implicit representations on implicit measures of attitude. Consequently, the effects of abstract supposition should be most apparent on self-reported measures and the effects of concrete learning on implicit measures.

One merit of drawing this novel distinction between concrete learning and abstract supposition is that it permits us to investigate whether Devine's (1989) dissociation model—in particular, its postulate that automatic attitudes generally show a greater resistance to change than their self-reported counterparts—still holds in a qualified sense, namely, when people are engaged in abstract supposition. A good reason for suspecting that this might be so is that egalitarian principles typically take the form of prescriptions about how one should think or feel (Devine, Monteith, Zuwerink, & Elliot, 1991). The "should" implies that one does not yet conform to an ideal that one consciously endorses. Yet, to consciously endorse an ideal, one must first be able to envisage it as a hypothetical state of affairs, with the assistance of suitably explicit cognition. If dual-process models are correct, then it follows plausibly that cognitive acts of this type should be relatively poor at purging the mind of more primitive and automatic biases (and of generating such primitive and automatic biases in the first place). In contrast, when people are engaged in concrete learning, one would make the opposite prediction. Here, automatic attitudes should be at least as malleable as their self-reported counterparts, given that concrete learning should affect the implicit memory system directly. Under such circumstances, one would *not* expect the predictions of Devine's (1989) dissociation model to be borne out: concrete learning should result in automatic attitudes being formed or changed at least as readily as their self-reported counterparts.[3]

With these theoretically derived implications as our initial springboard, we set out across four experiments, first to induce (Experiments 1 and 2), and thence to undo (Experiments 3 and 4), preferences for one imagined social group over another. In each study, participants were randomly assigned to conditions in which they either abstractly supposed or concretely learned in a variety of ways, that two social groups and their respective members possessed attributes of contrasting valence. The relative impact of these manipulations on self-reported and automatic attitudes was subsequently assessed with both rating scales and the IAT so that inferences about relative stability or malleability could be duly drawn.

In using imagined social groups rather than real ones, our research can be regarded as exemplifying the recent trend toward flexibly exploring the dynamics of attitude formation and change experimentally using novel attitude objects (De Houwer et al., 2001; Fazio, Eiser, & Shook, 2004; Glaser, 1999; Greenwald, Pickerell, & Farnham, 2002; Mitchell, 2004; C. J. Mitchell, Anderson, & Lovibond, 2003; Olson & Fazio, 2002). With regard to the current research question, our laboratory-based methodology has, despite its Spartan aspect, a number of distinct advantages. First, using unfamiliar stimuli as attitude objects maximizes the potential for new attitudes to take root (Cacioppo, Marshall-Goodell, Tassinary, & Petty, 1992). Second, by retaining control of the attitude-induction process, greater confidence is afforded that participants will construe attitude objects as intended and that those objects will correspond to self-reported and automatic attitudes (Fishbein & Ajzen, 1975). Third, by not focusing on real-life social groups, self-presentational biases that might otherwise complicate interpretation can be averted (Plant & Devine, 1998).

We were aware, of course, that by using imagined social groups, we ran the risk of compromising the realism of our investigation. To address this risk, we therefore included (in three of our four experiments) self-reported measures of attitude meaningfulness

[3] Two factors complicate this neat picture. First, explicit and implicit memory systems are unlikely to be completely distinct (Smith & DeCoster, 1999; Strack & Deutsch, 2004). Second, abstract supposition and concrete learning are unlikely to occur in the complete absence of one another. Thus, some limited cross-contamination—of concrete learning on explicit measures or of abstract supposition on implicit measures—is only to be expected. In addition, theory and praxis suggest that the degree of contamination is likely to be one-sided, with concrete learning affecting explicit measures more than abstract supposition affects implicit measures. The main reason is that it is nearly impossible for people to avoid activating explicit representations of objects about which they are knowingly encoding information (Gregg, 2003). This is eloquently attested to by the difficulties researchers have encountered in demonstrating the existence of learning in humans that bypasses consciousness entirely (De Houwer et al., 2001; Shanks & St. John, 1994; but see, for laudable attempts, Olson & Fazio, 2001; Petty & Jarvis, 1998). Moreover, given the common practice of tapping "implicit" constructs with measures that require self-reported evaluations (e.g., the Name Letter Task; Koole & Pelham, 2003), and of using self-report measures to assess the effects of "implicit" manipulations (e.g., Riketta & Dauenheimer, 2003), the effect of the implicit memory system on self-report measures would seem to be well-recognized by researchers. In contrast, there is no consensus as regards the degree of influence of the explicit memory system on implicit measures. Hence, we made this the primary focus of our article.

and as a precaution redid all pertinent analyses using only those participants scoring high on those measures.

Experiment 1

Our goal in Experiment 1, then, was to test the implication—plausibly derived from dual-process models and pertinent to the validity of Devine's (1989) dissociation model—that abstract supposition should have less of an impact on automatic preferences than on self-reported preferences. To do this, we used an attitude formation paradigm.

Method

Experiment 1 featured three conditions. In one condition (suppose) participants engaged in abstract supposition: They hypothetically assumed that one social group possessed positive traits and another one negative traits. In the two remaining conditions, participants engaged in concrete learning. They either (a) read a graphic story in which one group exemplified positive traits whereas another exemplified negative traits (narrative) or (b) repeatedly rehearsed positive trait associations toward one group and negative trait associations toward another (rehearsal). Our main hypothesis was that abstract supposition would induce more extreme self-reported preferences than either form of concrete learning, whereas either form of concrete learning would induce more extreme automatic preferences than abstract supposition.

We operationalized concrete learning in two different ways in order to achieve wider conceptual coverage of the construct. It could be argued that extracting pieces of information from a text (narrative) arguably qualifies as a more explicit form of learning than does rote rehearsal of paired associates (rehearsal). We were therefore open to the additional possibility that the former might prove superior to the latter at inducing self-reported preferences and the latter superior to the former at inducing automatic preferences.

Participants

Forty-six students from St. Anne's Convent School in Southampton, England, United Kingdom, participated in this study in return for funds allocated to the refurbishment of their recreational quarters. All participants were girls between the ages of 15 and 18, who were studying for A-levels.

Materials

Two novel social groups were created: *Niffites* and *Luupites*. Each group contained four members whose group membership could be readily identified from the spelling of their names. In particular, the names of Niffites featured a double consonant and ended in the syllable "nif" (e.g., Eska*nnif*), and the names of the Luupites featured a double vowel and ended in the syllable "lup" (e.g., N*ee*no*lup*). All names were pronounceable and contained three syllables. The names of both the groups and their members were constructed so as not to resemble any common word. The names finally chosen had been selected from a pool of similar candidates by virtue of being the most neutral and homogeneous in terms of their self-reported and automatic valence (Gregg, 2000).[4]

Induction of Group Preference

Before the induction proper, participants were asked to imagine that the two social groups described above actually existed. They were told that they would shortly be learning about what the groups were like and would be later asked questions about the groups. They were urged to keep clear in their minds for the entire duration of the study which group was which and which group possessed which characteristics. Participants were also familiarized with names of group members and with the orthographic distinction between them.

Narrative induction. Participants in the narrative condition read a graphic account of an intergroup conflict between the Niffites and Luupites. To encourage them to take this account seriously, they were informed that although a "real historical conflict" was being depicted, the identities of the groups involved were being concealed so that "perceptions of each group's character and behavior would not be biased by. . .knowledge of their true ethnic identity." One group, the Aggressors, was depicted as savage, ruthless, and brutal. The other group, the Victims, was depicted as civilized, accommodating, and constructive. The account of their conflict can be encapsulated as follows: The authoritarian Aggressors ran a military state from which the progressive Victims seceded. The Victims established a breakaway colony in a neighboring region and deservedly prospered. The Aggressors soon grew envious of the Victims" prosperity and invaded their territory; the Aggressors massacred the Victims in great numbers during the invasion and committed heinous crimes against them.

Rehearsal induction. Participants in the rehearsal condition completed a self-paced supraliminal priming task shown previously to produce effects on the IAT (Glaser, 1999). On each trial, group member names were briefly but visibly preceded by highly positive or negative trait adjectives (e.g., *benevolent*, *barbaric*). Each adjective appeared onscreen for 200 ms before being erased. After a 100-ms pause, the name of a group member then appeared in the same position. Participants' task was to indicate, as quickly as possible, the group to which a member belonged by pressing one of two keys. Labels for both groups remained on the upper portion of the screen throughout, each placed on the same side as the key used to classify group members. Throughout the entire priming task, negative trait adjectives primed the names of one group and positive trait adjectives the names of the other.

The induction consisted of a total of 240 trials divided into four blocks of 60. In each block, half the trials featured members of one group, half members of the other, intermixed at random. The sides of the screen on which the names of the two groups were placed alternated block by block. This was done to prevent the keys from being associated with positive or negative responses and to provide participants with pauses for rest. In addition, positive and negative adjectives were divided into two sets, their appearance in the induction alternating in tandem with the placement of group labels. Before the induction, participants completed eight practice trials on which they classified each of the eight group member names in the absence of primes.

Participants were informed before beginning the rehearsal induction that adjectives would be briefly flashed in advance of the names of the group members, that these adjectives would convey information about the character of those members, and that they should form impressions of the two groups on the basis of this information.

Supposition induction. Participants in the suppose condition were instructed to "suppose that the two groups [had] very different characters," that one was "very good. . .peaceful, civilized, benevolent, and law-abiding," whereas the other was "very bad. . .violent, savage, malicious, and lawless." Participants were also instructed to suppose, without any further elaboration, that the two groups consistently behaved in ways that justified these descriptions when they interacted with each other and with other groups.

Measures

Implicit measure. Automatic group preference was assessed using a five-block IAT (Greenwald et al., 1998). In each block, the words to be

[4] To this end, a series of pretests was run in which participants (a) explicitly rated the names of various groups and their members and (b) performed IATs featuring group names as labels and group members as stimuli (Gregg, 2000). Details are available from the first author.

classified appeared one after the other in the center of the screen. Category labels were displayed for the duration of the block on the upper left or right. Participants classified words by pressing a key on either side of the keyboard. Their task was to select, as quickly and accurately as possible, the key on the same side as the category label that corresponded to the presented word. If they did so correctly, the word disappeared; if they erred, a red X was flashed for 200 ms. Either way, the next word appeared 700 ms after each key press.

In Block 1, participants classified words of contrasting valence (*excellent*, *murder*) into categories of *Bad* and *Good*. In Block 2, they classified the names of group members (*Eskannif*, *Neenolup*) into the categories Niffite and Luupite. In Block 3, participants did a combination of both tasks. Blocks 4 and 5 were identical to Blocks 2 and 3, except that the category labels for Niffite and Luupite switched sides. Brief instructions were provided immediately before each block.

Each of the preparatory blocks (1, 2, and 4) comprised 13 trials and each of the critical blocks (3 and 5) comprised 49 trials. In the critical blocks—from which latency data were gathered—each of the four Niffite and four Luupite names was presented three times (24 trials) and each of the 12 Good and 12 Bad words was presented once (also 24 trials). Stimulus presentation was randomized under the constraint that the names (presented in black) and words (presented in blue) appeared on alternate trials. An additional word, randomly selected from one of the four stimulus groups, appeared on the first trial, but was ignored for the purposes of data analysis.

Depending on which group was portrayed positively and which negatively, the category label configuration in each of the critical blocks qualified as either compatible or incompatible. That is, if automatic preferences had been acquired in the manner intended by the manipulation, response latencies would have been accelerated in the compatible block but retarded in the incompatible block.

Explicit measure. On the basis of what they had previously "learned" about the two imagined social groups, participants were instructed to indicate what they now thought and felt about the groups. They duly rated each group, whose name was presented at random, on four bipolar scales featuring the following endpoints: *horrible–wonderful*, *unpleasant–pleasant*, *bad–good*, and *corrupt–virtuous*. Ratings were assigned by clicking on the appropriate onscreen digit, 1–7.

Postexperimental checks. At the end of the experiment, participants were asked two questions. The first question assessed the perceived realism of the paradigm: "How meaningful do you think it is to say that you held attitudes toward the Niffite and Luupite groups and their members?" Participants responded by clicking the appropriate number on a 7-point scale (1 = *Not at all*, 7 = *Extremely*). The second question assessed whether or not participants had correctly noted the intended valence of the two groups: "The information presented in this study attempted to convey which of the following impressions?" Participants responded by clicking one of two on-screen buttons, each containing an alternative answer: (a) "That Niffites are good and Luupites are bad" or (b) "That Luupites are good and Niffites are bad."

Procedure

After a preliminary orientation session, participants completed the entire study remotely on computer. To facilitate their participation, we set aside a quiet room on school premises that was equipped with 10 laptop computers. Over a 1-week period, participants were permitted to enter the room and complete the experiment during their free time. They activated the computer program running the experiment by clicking on a desktop icon. Participants were assigned to condition on the basis of personal identification numbers that they typed in. These numbers were printed on cards that participants had earlier drawn at random from a bag. Debriefing took place some weeks later in the context of a follow-up lecture given by Aidan P. Gregg to the participating students.

Results

Self-reported Preferences

Overall, participants rated the positively portrayed group more favorably than they rated the negatively portrayed group, $t(31) = 12.76$, $p < .0001$.[5] Moreover, this effect was only marginally moderated by condition, $F(2, 29) = 3.09$, $p = .061$, although the planned contrast between the suppose condition on the one hand, and the narrative and rehearsal conditions on the other, did reach significance, $t(29) = 2.48$, $p < .05$. In addition, participants' self-reported preferences were independently significant in all three conditions: $t_{\text{Narrative}}(8) = 6.25$, $p < .0001$; $t_{\text{Rehearsal}}(9) = 5.87$, $p < .0001$; $t_{\text{Suppose}}(12) = 11.31$, $p < .0001$ (see Figure 1, panel A). Thus, both abstract supposition and concrete learning proved capable of inducing self-reported preferences.

Automatic Preferences

Participants' IAT data were first screened to ensure that they met minimal standards of adequacy: an error rate of less than 25% and a suspicious latency rate (<150 ms or >5,000 ms) of less than 10%. All participants met these criteria. Subsequently, although trials on which errors were made (5.1%) were retained, trials yielding latencies either above 3,000 ms or below 300 ms (1.2%) were excluded as probable outliers. Remaining latencies were reciprocally transformed into speeds (speed =1,000/latency) before averaging (Figure 1, panel B).

Overall, participants responded more quickly in the compatible than in the incompatible block of the IAT, $t(31) = 5.15$, $p < .0001$. However, this effect was unmoderated by condition, $F(2, 29) = 1.16$, $p = .328$, and the planned contrast between the suppose condition on the one hand and the narrative and rehearsal conditions on the other did not reach significance, $t(29) = -.81$, $p = .426$. Moreover, participants' automatic preferences were independently significant in the rehearsal, $t(9) = 3.98$, $p < .005$, and suppose, $t(12) = 2.94$, $p = .012$, conditions and marginally significant in the narrative condition, $t(8) = 2.02$, $p = .078$ (see Figure 1, panel B). Thus, both abstract supposition and concrete learning proved capable of inducing automatic preferences.

Meaningfulness Check

Sixty-six percent of the sample rated the meaningfulness of their attitudes at or above the midpoint of the scale. Members of this subset accounted for 40% of participants in the rehearsal condition, 62% of participants in suppose condition, and 100% of participants in the narrative condition. The pattern of results was similar to that of the full sample, both in terms of their self-reported preferences,

[5] Fewer students participated than expected, and of those that did, 14 had to be excluded for confusing the intended valence of the groups. Low compliance rates may have reflected the absence of both direct supervision and individualized incentives. As a result, insufficient data were available to systematically assess the impact of three orthogonally counterbalanced methods factors (the valence of the social groups, the order of critical IAT blocks, and the order of explicit and implicit measures). Nonetheless, the results presented are a random sample of those that would have been obtained across all counterbalanced cells, and as such, provide an unbiased estimate of conditional differences.

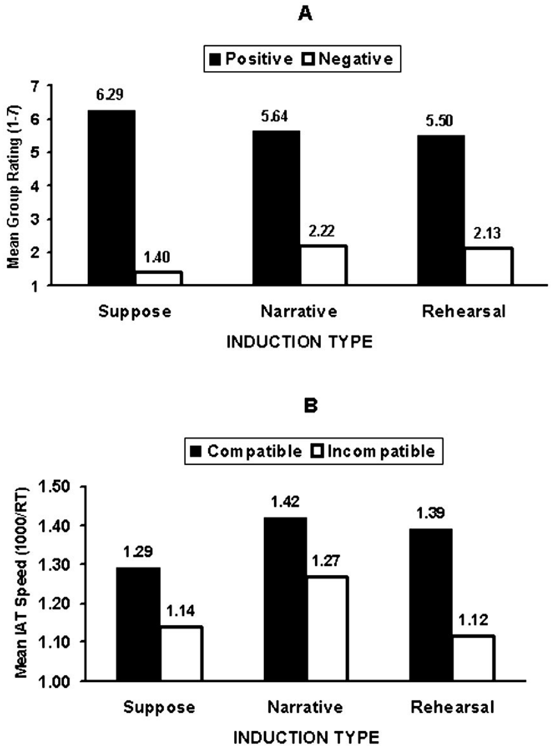

Figure 1. Experiment 1: Mean group ratings of self-reported preferences (A) and mean Implicit Association Test (IAT; Greenwald et al., 1998) response time (RT) for automatic preferences (B) for one imagined social group over another as a function of induction type. Panel A shows the results of participants rating each social group on the basis of what they had previously learned about them. Ratings were assigned by clicking on the appropriate onscreen digit, 1–7. Panel B shows the result of participants' assessment with a five-block IAT Depending on which group was portrayed positively and which negatively, the category label configuration in each of the critical blocks qualified as either compatible or incompatible. If automatic preferences had been acquired in the manner intended by the manipulation, response latencies would have been accelerated in the compatible block but retarded in the incompatible block.

$t_{Overall}(21) = 11.03$, $p < .0001$, $F_{Condition}(2, 19) = 6.63$, $p < .01$, and their automatic preferences, $t_{Overall}(21) = 3.14$, $p < .005$, $F_{Condition}(2, 19) < 1$.

Discussion

The results of Experiment 1 suggest that self-reported preferences can be readily induced by abstract supposition, and automatic preferences readily induced by concrete learning. However, despite suggesting that abstract supposition may be slightly more effective at inducing self-reported attitudes than concrete learning, the results do not suggest that concrete learning is any more effective at inducing automatic preferences than abstract supposition, whether the concrete learning in question involves reading a vivid narrative or repeatedly rehearsing associations. Consequently, the results are at variance with the implication that automatic attitudes should be relatively invulnerable to manipulations of explicit cognition. Rather, automatic attitudes seem to spring into existence on the basis of purely hypothetical assumptions.

Experiment 2

To ensure the unexpected findings of Experiment 1 were no fluke, we sought in Experiment 2 to replicate it conceptually on a larger sample. In the previous experiment, we had asked participants in the suppose condition to assume, without further elaboration, that two groups possessed a small number of evaluatively contrasting traits. The amount of information provided about the groups in this condition was, to say the least, scant compared with the other two conditions. Nonetheless, an objection could be raised that, because a few discrete items of information were indeed provided, concrete learning was not entirely absent from the suppose condition. To counter this objection, we adopted in Experiment 2 an even purer manipulation of abstract supposition: We tested whether participants could generalize automatic preferences previously acquired toward one set of stimuli to a new set of stimuli simply by hypothetically assuming that the two sets of stimuli were equivalent. Such a manipulation, we felt, would depend entirely on cognitions of a formal and symbolic nature.

Method

Participants

Seventy-three undergraduates from the University of Southampton, United Kingdom, participated in exchange for partial course credit. Of these, 5 were excluded for misunderstanding instructions. A computer malfunction resulted in the loss of data from 1 other participant.

Procedure and Design

All participants began by undergoing a procedure designed to induce self-reported and automatic preferences for one imagined social group over another. To ensure that these preferences were robustly acquired and regarded as subjectively meaningful, we used a double-barreled manipulation. This consisted of a sequential combination of the narrative and rehearsal induction procedures used in Experiment 1.

In addition, the two groups featured in the induction were not the Niffites and Luupites, but rather two equivalent groups, the *Jebbians* and *Haasians*. As before, the names of Jebbians featured a double consonant and ended in the syllable "jeb" (e.g., Fi*dd*ijeb), whereas the names of Haasians featured a double vowel and ended in the syllable "has" (e.g., R*oo*gihas). Also as before, all names could be pronounced, contained three syllables, and bore no obvious resemblance to any meaningful word. The stems "jeb" and "has" were chosen precisely because they did not overlap orthographically with "nif" and "lup," thereby permitting Jebbians, Haasians, Niffites, and Luupites to be visually distinguished from one another.

Following the preference induction procedure, participants were randomly assigned to one of two conditions. In the suppose condition, participants were asked to suppose that they had read the same narrative about and rehearsed the same associations toward two alternative groups. In particular, they were asked to hypothetically assume that the Jebbians and Haasians were equivalent to the Niffites and Luupites (or to the Luupites and Niffites). Care was taken to ensure that participants were clear on which of the two novel groups corresponded with which of the two original ones. The correspondence was asserted three times, and a direct instruction was given to participants to "take a moment to get these substitutions straight in your head." Participants were additionally familiarized with the names of individual Niffites and Luupites.

In the relearn condition, participants were put through another full preference induction procedure that featured Niffites and Luupites as stimuli. The only difference was that the phrase "Please read every sentence carefully" was added, capitalized and placed in parentheses, at the top of each page of the narrative. Its inclusion was designed to deter participants, who had read a version of the narrative earlier, from simply skipping through the new version.

To ensure that the results obtained were not affected by incidental variations of method, we orthogonally counterbalanced four factors: (a) the social group rendered preferable by the induction procedure, (b) the order in which critical IAT blocks appeared, (c) the order in which the explicit and implicit measures of group preference appeared, and (d) the mapping of old groups onto new groups (i.e., *either* [Niffite = Jebbian] + [Luupite = Haasian] *or* [Niffite = Haasian] + [Luupite = Jebbian]).

Measures

Implicit measure. To keep the experiment within a manageable time frame, we dropped the three preparatory blocks in the IAT. Instead, the 36 experimental trials of each critical block were preceded by 15 practice trials, and the two sets of trials separated by a single intermediate screen. Prior research indicates that strong and significant effects can be readily obtained even with such abbreviated IATs (e.g., Teachman, Gregg, & Woody, 2001). Across the 36 experimental trials, the same group member names and oppositely valenced words as in Experiment 1 were presented. At the end of each critical block, participants were given feedback about both their average latency and their overall error rate in order to motivate them to respond both quickly and accurately. Apart from these changes, and some superficial alterations in format, the IAT was essentially identical to that used in Experiment 1.

Explicit measure. Participants were instructed to rate their current feelings about the members of each group (1 = *bad*, 9 = *good*). The name of each member was presented twice at random, with the names of Niffites and Luupites alternating. Participants assigned ratings by pressing the appropriate numbered key.

Meaningfulness check. Participants read the following question on a postexperimental inquiry sheet: "How meaningful do you think it is to say that you held attitudes toward the Niffite and Luupite groups and their members?" They responded by circling the appropriate number on a 7-point scale (1 = *Completely meaningless*, 7 = *Completely meaningful).* On completion of the experiment, all participants were thanked, debriefed, and dismissed.

Results

Data Reduction

IAT data were reduced as in Experiment 1. Three participants were excluded for making excessive errors. Error and outlier rates for the final sample ($N = 64$) were, respectively, 5.5% and 1.6%. Statistical analyses of self-reported and automatic preferences took account of the four counterbalanced method factors.[6]

Self-reported Preferences

Overall, participants rated members of the positively portrayed group more favorably than they rated members of the negatively portrayed group, $F(1, 48) = 557.40$, $p < .0001$. Moreover, the effect was unmoderated by condition, $F(1, 32) < 1$, and independently significant in both conditions, $F_{Suppose}(1, 16) = 246.15$, $p < .0001$; $F_{Relearn}(1, 16) = 280.19$, $p < .0001$ (see Figure 2, panel A). Thus, both abstract supposition and concrete learning proved capable of inducing self-reported preferences.

Automatic Preferences

Overall, participants responded more quickly in the compatible than in the incompatible block of the IAT, $F(1, 48) = 46.80$, $p < .0001$. The effect was again unmoderated by condition, $F(1, 32) < 1$, and independently significant in both conditions, $F_{Suppose}$ (1, 16) = 17.15, $p < .001$, $F_{Relearn}(1, 16) = 26.98$, $p < .0001$ (see Figure 2, panel B). Thus, both abstract supposition and concrete learning proved capable of inducing automatic preferences of comparable magnitude.

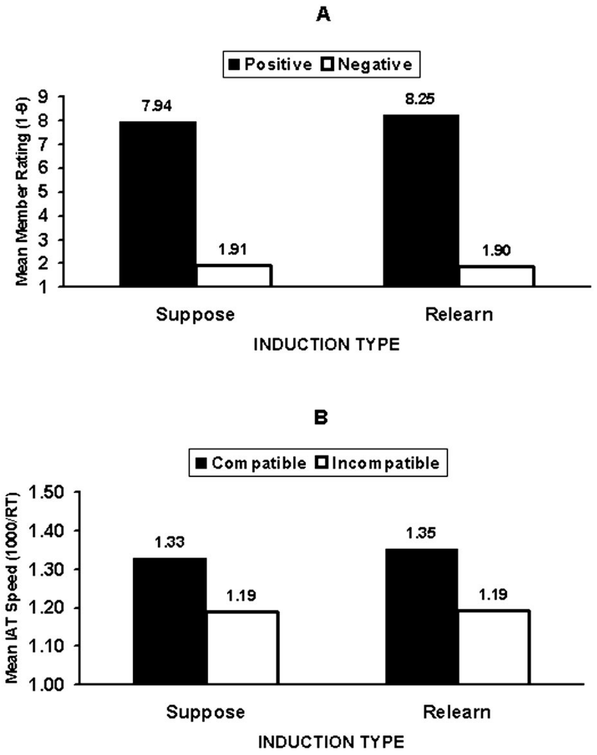

Figure 2. Experiment 2: Mean member ratings of self-reported preferences (A) and mean Implicit Association Test (IAT; Greenwald et al., 1998) response time (RT) of automatic preferences (B) for one imagined social group over another as a function of induction type. In the suppose condition, participants were asked to suppose that they had read the same narrative about, and rehearsed the same associations toward, two alternative groups. In the relearn condition, participants were put through another full preference induction procedure. The only difference was that the phrase "Please read every sentence carefully" was added to deter participants from who had read a version of the narrative earlier, from simply skipping through the new version. Panel A shows the results of participants rating their feelings about the members of each group (1 = *bad*, 9 = *good*). Panel B shows that, overall, the participants responded more quickly in the compatible than in the incompatible block of the IAT.

Meaningfulness Check

Meaningfulness data was supplied by 59 participants, of whom 78% gave ratings at or above the midpoint of the scale. This corresponded to 73% (22/30) of pertinent participants from the

[6] Because they are of only peripheral interest and would lengthen the article, the occasional higher-order interactions that occurred with method factors in this and other studies are not discussed. Full details are available from Aiden P. Gregg.

suppose condition and 83% (24/29) of pertinent participants from concrete condition. The results yielded by this subset were similar to those yielded by the full sample, both in terms of self-reported preferences, $F_{Overall}(1, 30) = 352.87$, $p < .0001$, $F_{Condition}(1, 16) < 1$, and automatic preferences, $F_{Overall}(1, 30) = 44.39$, $p < .0001$, $F_{Condition}(1, 16) = 1.44$, $p = .248$.

Discussion

The results of Experiment 2 confirm that automatic preferences, like self-reported ones, can be readily induced by abstract supposition as well as by concrete learning. In particular, automatic attitudes can be generalized from old objects to new objects simply by hypothetically assuming that both sets of objects are equivalent. Our first two experiments therefore empirically contradict what dual-process models can plausibly be taken as implying, namely, that automatic attitudes are relatively immune to sophisticated symbolic cognition.

Experiment 3

In Experiments 1 and 2, we had focused on attitude formation, in particular, on the induction of new self-reported and automatic preferences. In Experiments 3 and 4, we turned our attention to attitude *change*, in particular, to the undoing of recently induced self-reported and automatic preferences. We were eager to discover whether, because abstract supposition had earlier been sufficient to induce new automatic preferences, it would also be sufficient to undo established ones. Dual-process models, plausibly interpreted, implied that it would not. The findings of our previous two studies, in underscoring the acute malleability of automatic attitudes, suggested that it might. To put the matter to an empirical test, therefore, we induced self-reported and automatic preferences and then attempted to undo them using one or more relevant counterinductions.

Method

Experiment 3 featured two conditions. In one (suppose), participants were led to engage in abstract supposition by being given grounds to believe that the two groups they had previous learned about ought to have possessed each others" valences. In another (control), no such grounds were given: participants simply learned about the two groups and completed the dependent measures as usual. Participants" self-reported and automatic preferences were assessed twice, first after the original induction, and second after the counterinduction (or a filler procedure). This permitted us to test our predictions via shifts in group preference.

We expected that participants in the control condition would acquire and sustain self-reported and automatic preferences. That is, we expected no shift in group preference (except perhaps for a slight deterioration over time). Our expectations for participants in the suppose condition, however, were more tentative. On the one hand, we had interpreted classic dual-process models as implying that, although self-reported preferences would switch around, automatic preferences would remain directionally stable; abstractly supposing a counterfactual state of affairs would thus fail to override the impact of several minutes of prior concrete learning. On the other hand, our earlier findings had suggested that automatic preferences, just like self-reported ones, would indeed switch around. Having being induced by abstractly supposing that a hypothetical state of affairs obtained, automatic preferences could now be reversed by abstractly supposing that a counterfactual state of affairs did. Thus, the results of Experiment 3 would either confirm that automatic preferences could be shaped even by the most explicit of cognitions or would after all provide some evidence for the relative inertia of automatic preferences.

Participants

Fifty-one undergraduates attending Yale University were each paid $7 for their participation.

Procedure and Design

All participants began by undergoing the double-barreled preference induction procedure, which was followed by explicit and implicit measures of group preference. The critical manipulation (described below) came next, and was followed by a repeated set of explicit and implicit measures. All participants were then thanked, debriefed, and dismissed.

Participants in the suppose condition were given reason to suppose that the attitude induction procedure had been run the wrong way around. In particular, the experimenter told participants that she had just discovered a problem with the computer program. She claimed that although the program was intended to counterbalance, participant-by-participant, the valences assigned to the two groups, it in fact contained an error that made it portray one group as always positive and the other as always negative. The upshot, she alleged, was that some participants (including those she was addressing) had been exposed to portrayals of the groups that were diametrically opposite to those intended. To ensure that participants grasped this claim, the experimenter (a) began by stating the valence of the groups implied by the actual, and allegedly incorrect, induction procedure (e.g., "You probably noticed that the Niffites were described as bad and the Luupites as good. . ."); (b) proceeded to provide the explanation for the mix-up; and (c) ended by stating the valence that would have been implied by the hypothetical, and allegedly correct, induction procedure (e.g., ". . .so in fact the Niffites should have been described as good and the Luupites as bad."). Then, in a seemingly resourceful attempt to salvage some useful data, the experimenter suggested that participants might redo the categorization (implicit) and rating (explicit) tasks, this time supposing that the valences of the groups had been the other way around. She finally asked participants whether they clearly understood what they were being asked to do, and if no further clarification was requested, she stated one last time the new valences that the two groups were supposed to have and typed in a bogus five-digit code to activate the second set of measures.

Participants in the control condition were also informed, for reasons of parity, that there was a problem with the induction procedure. The experimenter told them that, owing to a clerical error, she had inadvertently entered a previously used participant number into the computer. The upshot, she alleged this time, was that the computer had failed to record any data, owing to an automatic safeguard that ensured no previously written file could be overwritten. She then asked participants if they would not mind redoing the categorization (implicit) and rating (explicit) tasks for her. Again, she typed in a bogus five-digit code to activate the second set of measures. (Note that, in both conditions, the error was attributed to the negligence of the principal investigator, to deflect criticism away from the experimenter herself).

Three method factors were orthogonally counterbalanced: (a) the social group rendered preferable by the induction procedure, (b) the order in which critical IAT blocks appeared, and (c) the order in which the explicit and implicit measures of group preference appeared. For each participant, the counterbalancing was maintained across both measurement occasions.

Measures

The explicit and implicit measures were identical to those in Experiment 2.

Results

Data Reduction

IAT data was reduced as in previous experiments. Three participants were excluded for making excessive errors. Error and outlier rates for the final sample ($N = 48$) were, respectively, 5.2% and 1.0% on the first IAT and 4.7% and 0.3% on the second. Statistical analyses of self-reported and automatic preferences took account of the three counterbalanced method factors.

Self-reported Preferences

After the preference induction procedure, participants rated members of the positively portrayed group more favorably than they rated members of the negatively portrayed group overall, $F_{Before}(1, 40) = 217.43$, $p < .0001$. This effect was also unmoderated by condition, $F(1, 32) < 1$, and was independently significant in both conditions: $F_{Suppose}(1, 16) = 113.09$, $p < .0001$; $F_{Control}(1, 16) = 114.00$, $p < .0001$. However, after the experimental manipulation, overall self-reported preferences were no longer apparent, $F_{After}(1, 40) < 1$. This was because they now differed significantly by condition, $F(1, 32) = 169.22$, $p < .0001$, and independently attained significance in opposite directions, $F_{Suppose}(1, 16) = 93.36$, $p < .0001$; $F_{Control}(1, 16) = 103.80$, $p < .0001$ (see Figure 3, panel A). Before–after analyses indicated that self-reported preferences shifted significantly in the suppose condition, $F_{Shift}(1, 16) = 119.70$, $p < .0001$, but not in the control condition, F_{Shift} $(1, 16) < 1$, and that the difference in the size of shifts was significant, $F(1, 32) = 99.62$, $p < .0001$.

In sum, participants in the suppose condition reversed their initial self-reported preferences after the experimental manipulation, whereas participants in the control condition retained their initial self-reported preferences.

Automatic Preferences

After the preference induction procedure, participants responded more quickly in the compatible block than in the incompatible block of the IAT overall, $F_{Before}(1, 40) = 41.34$, $p < .0001$. This effect was unmoderated by condition, $F(1, 32) < 1$, and independently significant in both conditions: $F_{Suppose}(1, 16) = 16.59$, $p < .001$; $F_{Control}(1, 16) = 24.75$, $p < .0001$. Moreover, after the experimental manipulation, overall automatic preferences remained apparent, $F_{After}(1, 40) = 24.70$, $p < .0001$. They were again unmoderated by condition, $F(1, 32) = 2.75$, $p = .107$, and independently attained significance in the same direction, $F_{Suppose}(1, 16) = 9.53$, $p = .007$; $F_{Control}(1, 16) = 17.43$, $p < .001$ (see Figure 3, panel B). Before–after analyses indicated that automatic preferences did not shift significantly in either the suppose condition, $F_{Shift}(1, 16) = 3.11$, $p = .097$, or in control condition, F_{Shift} $(1, 16) < 1$, and that the difference between the size of these shifts was nonsignificant, $F(1, 32) = 2.75$, $p = .107$.

In sum, participants in the suppose condition did not reverse their initial automatic preferences after the experimental manipulation. Instead, like participants in the control condition, they retained their initial automatic preferences.

Discussion

A comparison of the patterns obtained on explicit and implicit measures in Experiment 3 suggests that automatic preferences were less responsive to abstract supposition than self-reported preferences were.[7] In particular, when participants were given reasonable grounds to believe that their attitude toward the groups ought to have been the other way around, their self-reported preferences reversed, yet their automatic preferences remained directionally consistent with the thrust of the original induction. This finding is consistent with what dual-process models can be taken as plausibly implying: that (already established) automatic attitudes should be relatively resistant to change by abstract supposition. To ensure that this finding was robust, however, we attempted to conceptually replicate it in a final experiment.

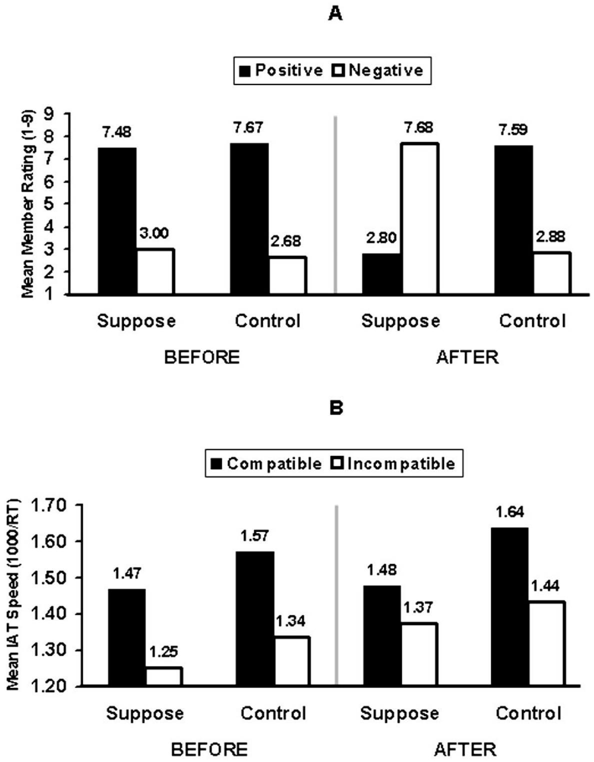

Figure 3. Experiment 3: Mean member ratings of self-reported preferences (A; scale, 1 = *bad*, 9 = *good*) and mean Implicit Association Test (IAT; Greenwald et al., 1998) response time (RT) of automatic preferences (B) for one imagined social group over another before and after different types of counterinduction. Panel A shows that participants in the suppose condition reversed their initial self-reported preferences after the experimental manipulation, whereas participants in the control condition retained their initial self-reported preferences. Panel B shows that after the preference induction procedure, participants responded more quickly in the compatible block than in the incompatible block of the IAT overall.

[7] Admittedly, it does not strictly follow that because the interaction for the explicit measure was significant and that for the implicit was not, that the difference between the two interactions was significant. However, because both measures were scaled along different metrics, had different reliabilities, and were influenced to different degrees by the manipulation, straightforward statistical comparison is precluded. Interpretations therefore rest on the strength and significance of directional effects observed for each measure. The same point applies to the subsequent experiment.

Experiment 4

We were mindful that the findings of Experiment 3 lay open to a possible objection. Might participants in the suppose condition have failed to change their minds about the valence of the groups? Might they have responded to the alleged mix-up by mechanically inverting their ratings? Might they have undergone no real shift in outlook but merely complied with experimental demands? Informal pretests and follow-up inquiries persuaded us that the answer in each case was no. The purpose of the cover story of the mix-up, after all, had been to provide participants with plausible grounds for reversing their preferences. Nonetheless, in the absence of formal checks of attitude meaningfulness, we could not objectively document the authenticity of participants' self-reports.

In Experiment 4, therefore, we adopted a tighter methodology. First, we modified the counterinduction procedure so that the grounds for reversing preferences were now based, not on an alleged experimental mix-up, but on new information provided about the two social groups, supplemented by arguments for why the groups might have reasonably switched characters. Second, we specifically asked participants, as we had in Experiments 1 and 2, how personally meaningful their attitudes toward the groups seemed to them.

We also addressed in Experiment 4 one possible explanation for why the findings of Experiment 3 might have conflicted with those of the first two experiments; that is, why abstract supposition was sufficient to induce, but not to undo, automatic preferences. In particular, we hypothesized that automatic attitudes might be generally easier to acquire than to eliminate. If this hypothesis were correct, then the relative inertia exhibited by automatic attitudes would be apparent, not when attempting to induce new ones (as we had found in Experiments 1 and 2) but only when attempting to undo existing ones (as we had found in Experiment 3). This, however, would in turn imply that existing automatic attitudes would tend to resist modification, not only by abstract supposition but also by other means, in particular, by concrete learning. In Experiment 4, we put this hypothesis to the test.

Method

We began, as in Experiment 3, by inducing group preferences in all participants. For reasons of time, however, we dispensed with the supraliminal priming task and relied solely on the narrative. Immediately afterward, we took explicit and implicit measures of group preference. We then randomly assigned participants to one of three conditions. In the suppose condition, participants were instructed to hypothetically assume, on the basis of very general information, that the characters of the two groups had switched over time. In the unlearn condition, participants were instructed to read a follow-up narrative describing in detail how characters of the two groups had switched over time. In the control condition, participants were instructed to read a neutral narrative describing in detail the flora and fauna indigenous to the geographical region inhabited by the groups. We then retook explicit and implicit measures of group preference.

We hypothesized that automatic preferences, relative to self-reported preferences, would be generally easier to acquire than to eliminate. Thus, we predicted that self-reported preferences would, in the suppose and unlearn conditions, switch from pre- to postmanipulation, whereas automatic preferences, in those same conditions, would persist from pre- to postmanipulation. In contrast, we predicted that both self-reported and automatic preferences would, in the control condition, persist from pre- to postmanipulation.

Participants

Forty-eight undergraduates attending the University of Würzberg participated. The experiment itself, which lasted about 30 min, began 30 min after participants had completed some other unrelated experiments. When the experiment concluded, participants went on to complete further unrelated experiments for 30 more min. At the end of the entire session, participants were debriefed, thanked, and paid €9 (at the time, about $9).

Materials

The translation of stimulus materials into German was carried out by Beate Seibt, who is a native speaker. Narrative information was provided in printed booklets in order to present particular German characters in their most familiar form (e.g., β). A computer program presenting transliterated characters (e.g., *ss*) provided general instructions for the experiment and administered explicit and implicit measures of group preference.

Procedure and Design

The induction narrative, read by all participants, was essentially identical to that used in all previous experiments. What participants read thereafter, however, differed depending on the experimental condition to which they had been assigned.

In the suppose condition, participants were instructed to suppose that "the two groups switched their characters, so that the good one became bad, and the bad one became good" and to keep supposing that this was the case "until the very end of this experiment, while [doing] each of the remaining tasks." To make this supposition seem more plausible, participants were informed that the massacre described in the induction narrative had two paradoxical effects. First, it "awakened the [Aggressors] to the immorality of their society and acted as a stimulus for reform," with the result that the Aggressors eventually became "peaceful, civilized, benevolent, and law-abiding." Second, it "destabilized and embittered the [Victims], leading them to mount terrorist counterstrikes against the [Aggressors]," with the result that the Victims eventually became "violent, savage, malicious, and lawless."

In the concrete condition, participants read a narrative replete with vivid and specific detail about how the two groups switched characters. To maximize the likelihood of reversing participants' preferences, we extended this narrative so that it exceeded the original in length. Its content can be encapsulated as follows. Aggressor society was already crumbling under the weight of its own corruption and iniquity; the massacre of the Victims proved a turning point; progressive elements within Aggressor society successfully mobilized and toppled the military dictatorship. With the rule of law and democratic government now established, its economy revived and culture flourished; the newly idealistic Aggressors sought to make amends for past wrongs committed against the Victims. Unfortunately, Victim society, always fragile, had been irreparably shattered; rival factions among the Victims, struggling to survive, fought against one another, and civil society fell apart. Bitter hatred of the Aggressors persisted; seeking retribution, the Victims waged a terrorist campaign against the Aggressors and consistently exploited peaceful Aggressor overtures by murdering innocent Aggressor civilians. Ultimately, a state visit by the Aggressor president, a brave and noble gesture aimed at resolving the conflict, ended in catastrophe when Victims assassinated him, plunging the region into fresh chaos and terror.

In the control condition, participants read an account of various flora and fauna native to the geographical reaction occupied by both Aggressors and Victims. The length of this account matched that of the counterinduction used in the unlearn condition. Three method factors (preferable group, IAT block, preference measure) were again consistently counterbalanced across measurement occasions.

Measures

The explicit and implicit measures of group preference and the one-item measure of attitude meaningfulness (also administered twice), were identical to those featured in Experiments 1 and 2.

Results

Data Reduction

IAT data were reduced as in previous experiments. All participants met inclusion criteria. Error and outlier rates were, respectively, 3.5% and 1.3% on the first IAT and 0.5% and 0.3% on the second. Statistical analyses of self-reported and automatic preferences took into account the three counterbalanced method factors.

Self-reported Preferences

After the induction procedure, participants rated the positively portrayed group more favorably than they rated the negatively portrayed group overall, $F_{Before}(1, 40) = 497.11$, $p < .0001$. This effect was unmoderated by condition, $F(1, 24) < 1$, and was independently significant in all three conditions: $F_{Suppose}(1, 8) = 148.63$, $p < .0001$; $F_{Unlearn}(1, 8) = 410.61$, $p < .0001$; $F_{Control}(1, 8) = 124.04$, $p < .0001$. However, following the experimental manipulation, self-reported preferences differed significantly by condition, $F(1, 24) = 17.27$, $p < .0001$ (see Figure 4, panel A). In particular, those participants in the control condition were significantly different from, as well as directionally inconsistent with, those participants in the suppose and unlearn conditions, who did not differ reliably by Tukey's honestly significant difference (HSD) test at $p < .05$. Self-reported preferences were also independently significant in all three postmanipulation conditions, $F_{Suppose}(1, 8) = 5.67$, $p = .045$; $F_{Unlearn}(1, 8) = 11.37$, $p = .01$; $F_{Control}(1, 8) = 16.95$, $p = .003$. Before–after analyses indicated that self-reported preferences shifted significantly in the suppose condition, $F_{Shift}(1, 8) = 106.75$, $p < .0001$, and the unlearn condition, $F_{Shift}(1, 8) = 117.43$, $p < .0001$, but not in the control condition, F_{Shift} $(1, 8) = 2.32$, $p = .133$. Moreover, the size of the shift differed reliably by condition, $F(2, 24) = 24.73$, $p < .0001$, with control participants undergoing a significantly smaller shift than suppose and unlearn participants (by Tukey's HSD at $p < .05$).

In sum, participants in the suppose and unlearn conditions reversed their initial self-reported preferences following the experimental manipulation, albeit not symmetrically. Participants in the control condition, in contrast, retained their initial self-reported preferences.

Automatic Preferences

After the induction procedure, participants responded more quickly in the compatible block than in the incompatible block of the IAT overall, F_{Before} $(1, 40) = 19.44$, $p = .0001$. Moreover, the effect was unmoderated by condition, $F(1, 24) < 1$, and was independently significant in the suppose condition, $F_{Suppose}(1, 8) = 14.14$, $p < .006$, as well as marginally significant in the unlearn and control conditions: $F_{Unlearn}(1, 8) = 3.81$, $p < .087$; $F_{Control}(1, 8) = 3.73$, $p = .09$. Moreover, after the experimental manipulation, automatic preferences remained unmoderated by condition, $F(1, 24) = 1.22$, $p = .312$, and directionally consistent (see Figure 4, panel B). Nonetheless, automatic preferences fell short of significance in the suppose condition, $F_{Suppose}(1, 8) = 2.71$, $p = .139$, despite independently attaining it in the unlearn and control conditions: $F_{Unlearn}(1, 8) = 8.20$, $p = .021$; $F_{Control}(1, 8) = 6.95$, $p = .03$. Before–after analyses indicated that automatic preferences shifted significantly in the suppose condition, $F_{Shift}(1, 8) = 6.25$, $p < .037$, but not in the unlearn or control conditions, both $F_{Shifts}(1, 8) < 1$. In addition, the size of the shift differed reliably by condition, $F(2, 44) = 5.35$, $p = .012$, with suppose participants undergoing a significantly larger shift than control participants (by Tukey's HSD at $p < .05$).

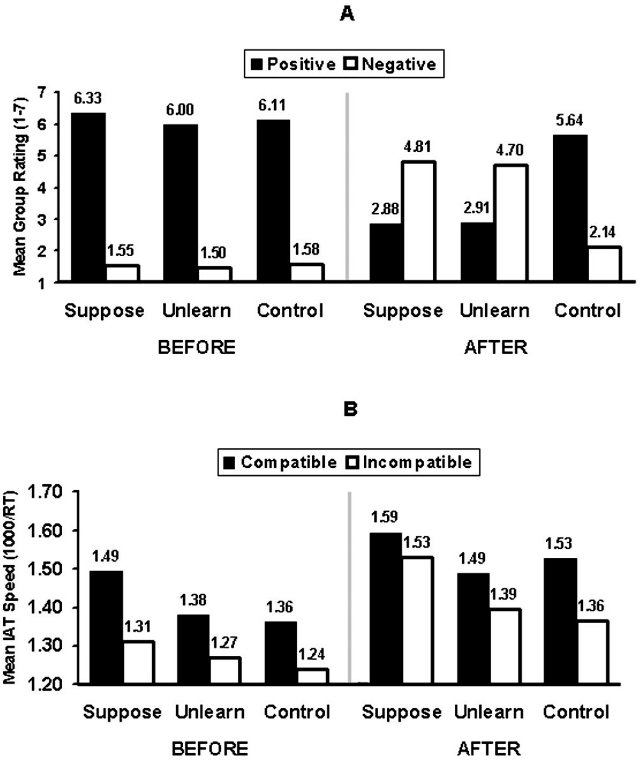

Figure 4. Experiment 4: Mean group rating of self-reported preferences (A, scale of 1–7) and mean Implicit Association Test (IAT; Greenwald et al., 1998) response time (RT) of automatic preferences (B) for one imagined social group over another before and after different types of counter-induction. In the suppose condition, participants were instructed to assume that the characters of the two groups had switched. In the unlearn condition, participants were instructed to read a follow-up narrative describing how the characters of the two groups had switched. In the control condition, participants were instructed to read a neutral narrative. Panel A shows participants rated the positively portrayed group more favorably than they rated the negatively portrayed group overall. Before–after analyses indicated that self-reported preferences shifted significantly in the suppose and unlearn conditions but not in the control condition. Panel B shows that after the induction procedure, participants responded more quickly in the compatible block than in the incompatible block of the IAT overall.

In sum, participants in the suppose and unlearn conditions did not reverse their initial automatic preferences following the experimental manipulation. However, the findings suggested that the automatic preferences of suppose participants became less extreme.

Meaningfulness Check

After the first assessment, 85% of the sample gave meaningfulness ratings at or above the midpoint of the scale. This subset comprised 100% of suppose participants, 81% of unlearn participants, and 75% of control participants. Following the second assessment, 65% of the sample gave meaningfulness ratings at or above the midpoint of the scale. This subset comprised the same percentages of suppose and unlearn participants but only 44% of the control participants. The marked decrease in the control condition seems to have been the inadvertent consequence of drawing participants' attention away from the two social groups. The fact that perceived meaningfulness was unaffected in the suppose and unlearn conditions, $t(32) = 1.31$, $p = .199$, suggests that the counterinductions they contained were regarded as credible.

The pattern of results obtained for participants who consistently regarded their attitudes toward the groups as relatively meaningful was similar to that obtained for the full sample. Their postinduction self-reported preferences were significant overall, $F_{Before}(1, 26) = 364.25$, $p < .0001$, and unmoderated by condition $F(1, 14) < 1$. Their postmanipulation self-reported preferences, however, were moderated by condition, $F_{After}(1, 26) = 7.48$, $p < .0001$, with those of control participants differing reliably from, and pointing in the opposite direction to, those of suppose and unlearn participants (by Tukey's HSD at $p < .05$). Shifts in self-reported preference were also moderated by condition, $F(2, 14) = 10.41$, $p < .0001$, with control participants undergoing a reliably smaller shift (by Tukey's HSD at $p < .05$).

In addition, their postinduction automatic preferences were also significant overall, $F(1, 26) = 13.97$, $p < .001$, and unmoderated by condition $F(1, 14) < 1$. Moreover, their postmanipulation automatic preferences were only marginally moderated by condition, $F(1, 14) = 3.15$, $p = .079$, with post hoc tests confirming that they did not differ reliably (by Tukey's HSD at $p < .05$). Shifts in automatic preference were nonetheless significantly moderated by condition, $F(2, 14) = 5.97$, $p = .013$, with suppose participants undergoing a reliably larger shift than either suppose and unlearn participants (by Tukey's HSD at $p < .05$).

Discussion

After the experimental manipulation, the self-reported preferences of participants in the suppose and unlearn condition switched around to become significant in the opposite direction. However, their automatic preferences clearly did not follow suit. Instead, they remained directionally consistent with the thrust of the original induction procedure. In other words, *both* abstract supposition and concrete learning failed to undo participants' automatic preferences to the same degree that they undid their self-reported preferences. This suggests that automatic preferences, once established, are generally difficult to dislodge, with concrete learning proving as ineffectual in dislodging them as abstract supposition is. Our new hypothesis concerning the relative stability of implicit preferences was supported.

Nonetheless, depending on which analyses are highlighted, it could be argued that Experiment 4 also yielded limited evidence for the malleability of existing automatic preferences. Within the suppose condition group, automatic preferences declined significantly from pre- to postmanipulation; the degree of decline significantly exceeded that observed in the control condition group; and the postmanipulation effect itself fell somewhat short of significance. These effects suggest that existing automatic preferences are not wholly immune to the perturbing impact of abstract supposition. Nevertheless, the magnitude of postmanipulation automatic preferences did not differ significantly by condition. The big picture is one of consistency, not change. Moreover, within the unlearn condition group—in which participants spent several minutes reading vivid details about how the characters of the groups had been transformed—no evidence of any decline in their automatic preferences emerged. Instead, their original automatic preferences became nonsignificantly more pronounced.

These findings are worth considering along with those recently reported by Foroni and Mayr (in press). Here, participants successively encountered, in counterbalanced order, (a) stereotype-confirming information indicating that flowers and insects are, respectively, beneficial and noxious and (b) stereotype-challenging information indicating that flowers and insects are, respectively, noxious and beneficial. In addition, participants either encountered this information (a) in the context of a science-fiction scenario or (b) in the form of a simple instruction to believe it. Only when the counterstereotypical information was embedded in a science-fiction scenario did it exert any impact on automatic preferences; no impact was observed when it was conveyed as a simple instruction. Additionally (although the authors do not comment on the fact) exposure to the counterattitudinal scenario plainly did not reverse participants" automatic preferences but merely attenuated them. Thus, Foroni and Mayr found, in effect, that concrete learning had a limited impact on existing automatic preferences but that abstract supposition had none, whereas we found in Experiment 4 that concrete learning had no impact on existing automatic preferences but that abstract supposition may have had some.

Our interpretation is that both sets of findings, considered together, support the thesis that automatic preferences are relatively resistant to change. Neither in Experiment 4 nor in Foroni and Mayr's research did existing automatic preferences reverse direction. In Experiment 4, it is particularly notable that they did not, given that corresponding self-reported preferences did. In Foroni and Mayr's research, it is not quite so telling, given that the scenario was hardly meant to extend to all flower and insects without exception (and in any case, no self-reported measures of preferences were taken). Nonetheless, given the evident ease with which automatic preferences can be created (our Experiments 1 and 2), one might have been forgiven for expecting that participants" automatic preferences would have been dramatically modified by what they were consciously contemplating. However, this simply did not happen: the changes to automatic preferences that did occur were at best modest and averaging across both experiments, there is little reason to believe that concrete learning or abstract supposition was a more powerful agent of change in this regard. Note that we are *not* asserting that automatic attitudes, once formed, can *never* change in a substantial way; obviously they can, depending on the magnitude and extent of the relearning involved. Indeed, such relearning has been successfully effected both in experiments (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000) and in quasi-experiments (Rudman, Ashmore, & Gary,

2001) that feature concrete learning. Rather, we suggest that, in view of the results of all four experiments, automatic preferences are more difficult to undo than to induce relative to their self-reported counterparts.

General Discussion

Overview of Findings

Across four experiments, we assessed the relative malleability of self-reported and automatic attitudes by attempting to induce and undo preferences for one imagined social group over another.

In Experiment 1, we found that abstractly supposing that one group was good and another bad was sufficient not only to induce new self-reported preferences but also to induce new automatic preferences. Moreover, the automatic preferences induced by abstract supposition did not differ significantly from those induced by concrete learning (whether it involved reading a narrative or rehearsing multiple associations).

In Experiment 2, we conceptually replicated the effect using a generalization paradigm. We found that abstractly supposing that two new groups were equivalent to two old groups was sufficient to transfer automatic preferences from one group pair to the other. Indeed, the automatic preferences so transferred were statistically indistinguishable from those induced by fresh concrete learning (a combination of reading a narrative and rehearsing associations). Automatic attitudes again proved to be as malleable as their self-reported counterparts.

However, when we moved from studying attitude formation to studying attitude change, a different picture emerged. Specifically, we found in Experiment 3 that abstractly supposing that the characters of two groups had been assigned the wrong way around failed to reverse the automatic preferences that had recently been acquired toward those groups, although it did succeed in reversing self-reported preferences toward them. In Experiment 4, we furthermore found that when participants not only abstractly supposed but also concretely learned that the characters of two groups had changed over time, their automatic preferences failed to reverse (although abstract supposition may have had some impact).

Devine's (1989) dissociation model predicts, among other things, that automatic attitudes will be generally harder to shift that their self-reported counterparts. In an effort to identify some boundary conditions under which this prediction might be true, we began by plausibly interpreting dual-process models as implying (among other things) that automatic attitudes would be more responsive to concrete learning than to abstract supposition. However, our cumulative findings then strongly suggested that we had misidentified the relevant boundary conditions. Experiments 1 and 2 established that automatic preferences, like self-reported ones, could be readily induced, not only through concrete learning but also through abstract supposition, a form of highly explicit, symbolic cognition. Although Experiments 3 and 4 established that recently acquired automatic preferences resisted attempts to undo them, concrete learning did not emerge as a more effective means of undoing them than abstract supposition did. It is remarkable that despite considerable theoretical precedent, concrete learning never exerted a larger impact on automatic attitudes than abstract supposition did.

Because we began by putting forward a theoretically derived hypothesis and calling its viability into question on the basis of experimental data, it behooves us to listen carefully to what that data has been trying tell us and to draw together plausibly the various strands of evidence. The most parsimonious inductive explanation for our cumulative findings, we contend, is that automatic attitudes are *asymmetrically malleable*. That is, like credit card debt and excess calories, they are easier to acquire than they are to cast aside. Thus, when people construe an object for the first time, their conscious fondness or antipathy for it is swiftly supplemented by an automatic positive or negative reaction. However, once people have acquired an attitude toward the object, attempts to subsequently undo it are differentially successful at different levels of the mind and lead its automatic component to lag behind its conscious one. Thus, Devine's (1989) key prediction—that automatic attitudes will be generally be harder to shift that their self-reported counterparts—may be correct after all, not under the boundary conditions that we initially proposed but under a new set of boundary conditions that our data have subsequently suggested.

If the postulated dynamic seems odd, then consider by way of analogy the nebulous images often presented in visual perception textbooks. Consisting only of fragments of the original object depicted, these images typically take perceivers a moment or two to consciously interpret. However, once a conscious interpretation is arrived at, it sticks: that is, it is immediately and automatically activated upon subsequent exposure to the nebulous image. Indeed, it becomes essentially impossible for perceivers not to make the appropriate interpretation, even years after having "seen" the image. Take a look at Figure 5. After a period of head-scratching, you are likely to come to see it as representing a Dalmatian dog, head bowed away from the viewer. The next time you open the relevant page of this journal, however, you will instantly recognize it as such. Indeed, you will be perceptually incapable of *not* recognizing it as such. We contend that automatic attitudes operate like rapidly established perceptual defaults: although they can initially be engendered by conscious cognition, they later become relatively resilient to its influence.

Our thesis is also consistent with intriguing research at the interface of decision making and neurobiology. Several experiments by Bechara and his colleagues (Bechara, A. R. Damasio, H. Damasio, & Anderson, 1994; Bechara, Tranel, & Damasio, 2000; see also Damasio, 1994) have vividly illustrated what happens when automatic preferences abnormally fail to stick. Participants

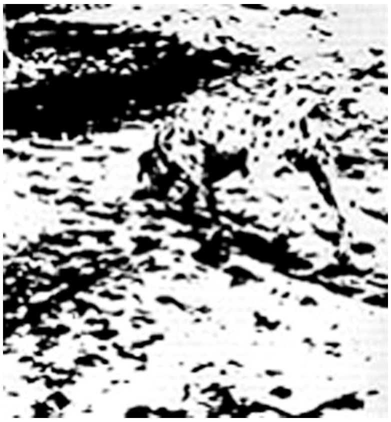

Figure 5. Nebulous image (Clue: It's No. 102).

play a simulated gambling game designed to mimic the vicissitudes of everyday life. On each turn, they choose a card voluntarily from one of four decks. The chosen card then specifies whether participants get paid or have to pay out. Most of the time, matters are arranged so that cards from two decks yield greater average gains (or smaller average losses) than cards from the other two decks. Every so often, however, a wildcard from one of the two normally lucrative decks leads to a huge loss (or a wildcard from the two normally unprofitable decks leads to a huge gain). The performance of patients with damage to the ventromedial prefrontal cortex is then compared against that of suitable controls. It turns out that controls, after a period of trial and error, develop an enduring negative bias against the deck "tainted" by the occasional huge loss (or an enduring positive bias toward the deck "blessed" by the occasional huge gain). However, the prefrontal patients do not: their aversion (or attraction) to the deck containing the wildcard quickly wanes, with the result that, in the long run, they incur substantial losses (or fail to secure substantial gains).

These findings are particularly interesting given that ventromedial prefrontal patients otherwise function at a very high cognitive level, preserving all their powers of rational analysis and verbal articulation. What they appear to lack is any facility for quickly developing a long-lasting dislike of, or liking for, options that, every once in a while, prove catastrophic or providential. In the absence of such ingrained, visceral biases—in effect, automatic preferences for one deck or another—the quality of their decision making suffers. For all their "explicit" acumen, they fall prey to imprudence and myopia.

These findings suggest that implicit cognition serves as a kind of stabilizing anchor that keeps the sometimes wayward sail of explicit cognition in check. In a social environment in which habitually good options can occasionally turn pear-shaped (e.g., a romantic partner cheats), or habitually bad options can occasionally turn golden (e.g., working overtime leads to promotion), slow-to-change defaults that are quickly established but thereafter difficult to eliminate (e.g., persistent antipathy to the partner, redoubled commitment to the job) are liable to pay off in the long run (e.g., a more faithful partner being sought out, a more successful career being achieved). Implicit cognition, together with its neural underpinning, may guide long-term decision making in the real world by acting as a hidden conservative counterweight, reigning in the liberal flexibility of explicit cognition. Indeed, interesting correlations have already been established between performance on implicit measures and certain patterns of neural activation (Phelps et al., 2000; Phelps, Cannistraci, & Cunningham, 2003).

Reconciling Divergent Findings

One important issue remains to be addressed: How might we reconcile our finding that established implicit attitudes are relatively stable with other published literature suggesting that they are relatively malleable (e.g., Blair et al., 2001; Dasgupta & Greenwald, 2001; Lowery et al., 2001; J. P. Mitchell et al., 2003)? We suggest that there are four primary reasons for the discrepancy.

First, there is the issue of what counts as "malleable" or "stable." Do these terms correspond, respectively, to shifts in automatic attitude that attain, or that fail to attain, statistical significance? Or do they mean, most substantially, shifts in automatic attitude that also meet or that fail to meet additional criteria, such as a manifest change in direction or a manifest change in concordance with self-reported attitude? If statistical significance is used as the sole criterion of malleability, then, by the conventional logic of hypothesis testing, only evidence for malleability can emerge, because the alternative would be a null result whose interpretation must remain equivocal (cf. Krueger & Funder, 2004). Furthermore, given that the metric of psychological variables is typically assumed rather than established (Blanton, Jaccard, & Gonzales, 2004; Judd & McClelland, 1998), it is difficult to determine whether any observed shift in automatic attitude of any particular magnitude is meaningful or not, when considered in isolation. In contrast, if independent criteria for meaningfulness are used—like directional change and external concordance—then interpretation is facilitated. For example, if, in Experiments 3 and 4, automatic attitudes had been as malleable as self-reported attitudes, then they should have switched around with them; that is, they should have (a) changed direction and (b) remained concordant with self-reported attitudes. However, this did not happen. Instead, automatic attitudes (a) remained directionally consistent and (b) changed from being concordant with their self-reported counterparts to being discordant with them. This provided powerful support for the relative stability of automatic preferences.

We suggest that other studies may have found evidence for the malleability of automatic preferences in virtue of relying primarily on statistically significant shifts as a criterion of malleability. Although such evidence, especially when considered in the aggregate, undeniably carries weight, it is not, for the reasons stated above, as telling as the evidence that is also based on independent criteria. Attempting to satisfy these independent criteria sets the evidential bar higher. For example, if we had drawn inferences about the malleability of automatic preferences solely on the basis of statistical significance, we might have casually concluded in Experiment 4 that existing automatic preferences are (according to one analysis) "malleable" in the face of abstract supposition. However, this conclusion would only have been warranted to the extent that we meant here "not utterly impervious to influence." If we meant by something more substantial, such as "readily amenable to radical reversal" or "prone to change as much as self-reported preferences," then we plainly could not have drawn the conclusion. In fact, we found consistent evidence in Experiments 3 and 4 that existing automatic preferences were not malleable in this more substantial sense.

The second reason for the discrepancy between our results and the published literature is that the latter has so far largely addressed the malleability of automatic attitudes toward traditionally stigmatized members of society. Although this research spotlight is clearly important, it has one key drawback when it comes to determining the relative malleability of self-reported and automatic preferences: The attitudes under investigation are inherently reactive. Consequently, the self-reported attitudes that participants express—especially in a liberal university environment—are liable not only to reflect what participants truly think and feel but also how they believe others wish them to think and feel, as well as how they personally believe they ought to think and feel (Plant & Devine, 1998). Hence, self-reported attitudes are likely to vary less than they should: the invalidating influences of self-presentation and self-deception exert an artificial stabilizing influence. This may explain, for example, why Dasgupta and Greenwald (2001) found that although automatic attitudes shifted after exposure to racial exemplars, self-reported attitudes did not. In our experiments, however, it is highly unlikely that any such invalidating influences were operative. Our participants' self-reported attitudes were free to vary with what the participants spontaneously felt and thought. Rather, the danger for us was that the validity of our self-reported attitudes might be curtailed by the lesser realism of our paradigm. Conscious of this danger, however, we took pains to assess empirically the perceived meaningfulness of the attitudes that we induced and undid. Our results showed that most participants regarded their attitudes as meaningful and that the same patterns of results were obtained for these participants as for all participants.

A third possible reason for the discrepancy is that, in our counterinduction manipulations, participants were instructed to contemplate either (a) how the characters of social groups had changed from what they had previously been (Experiment 4) or (b) how the attitudes that they themselves held toward the groups should have changed (Experiment 3). In contrast, the manipulations used by other researchers have not directed participants to contemplate attitude change directly. Instead, they have simply involved presenting participants with exemplars of the social groups under consideration, so that the groups themselves remain the exclusive objects of attention. It could be that exemplar-based manipulations, being more centered on the attitude objects themselves, are better suited to shifting automatic attitudes, given that change is perhaps a more cognitively complex notion than mere exemplification.[8]

The fourth possible reason for the discrepancy between our results and the published literature—and arguably the most theoretically significant— derives from the fact that attitudes toward stigmatized social groups, having been acquired in the real world, are likely to be multidimensional and polyvalent, whereas attitudes toward Niffites and Luupites, having been acquired in the laboratory, are likely to be unidimensional and univalent. This means that one method of attitude change, focus-switching, is more likely to occur in the former case than in the latter. Consider a person who, overall, has a negative automatic attitude toward Blacks. Their automatic attitude may reflect the integration of a large number of subsidiary automatic attitudes, the majority of which are negative, but a minority of which are positive (Anderson, 1981; Devine & Baker, 1991; Kunda & Thagard, 1996). For example, the person may possess a smaller number of positive automatic attitudes toward high-status Blacks while possessing a greater number of negative automatic attitudes toward low-status Blacks. This would make it possible for an experimental manipulation to increase the relative accessibility of the positive subtype and thereby increase the overall favorability of the automatic attitudes toward Blacks exhibited on a particular occasion.

Note that such a mechanism is sufficient to explain why recent exposure to a positive exemplar of a social category substantially reduces the negativity of automatic attitudes expressed toward that category (Lowery et al., 2001; Dasgupta & Greenwald, 2001), as well as the debiasing effects of bringing vivid counterstereotypical exemplars to mind (Blair et al., 2001). Indeed, contextual variations designed to render accessible subtypes of varying valence have been shown experimentally to moderate the valence of automatic racial attitudes (Wittenbrink, Judd, & Park, 2001). Moreover, despite initial indications to the contrary (De Houwer et al., 2001), it would appear results on the IAT and Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001) are a function, not only of the categories featured but also of the items classified under them, which influence how those categories are interpreted (Govan & Williams, 2004; J. P. Mitchell, et al., 2003). This again suggests that everyday social targets enjoy considerable latitude of interpretation, one that can be exploited by attitude change manipulations in applied settings. In contrast, the automatic attitudes that participants hold toward novel stimuli such as Niffites and Luupites are likely to be primarily if not solely informed by the information and instructions presented over the course of the experiment. Hence, no preexisting subtypes would be readily available in memory on the basis of which participants might construct alternative automatic attitudes. Any subtypes participants used would have to be generated by extrapolation from other similar instances, a feasible but more difficult, proposition.

That said, attitudes toward everyday targets are not *necessarily* multivalent: there are occasional exceptions. For example, many people would be unable to credit the Nazis with any redeeming features, whereas many others would be unable to regard their God as possessing any imperfections. Furthermore, targets with which people have only recently become acquainted are more liable to be associated, in the beginning, with information of a consistent valence. Thus, the results of our experiments may even generalize directly to a subset of attitudes toward everyday targets.

The most critical conceptual question that arises is whether focus-switching can be regarded as genuine attitude change (cf. Devine, 2001, p.759). It might be argued that, unless the exemplars underlying attitude prototypes or subtypes are replaced by other exemplars, any positive shifts observed will be inevitably superficial, merely reflecting an alteration in exemplar *accessibility* rather than in exemplar *availability*. A legitimate concern is that such shifts in accessibility might prove ephemeral—temporary departures from a more negative and enduring default attitude. (An analogy would be the temperature of an air-conditioned room, which, though it may be momentarily perturbed by external conditions, tends to revert to its thermostatic set point.) If so, then tests of attitude change using multivalent real-life targets might overestimate the magnitude of true attitude change, whereas artificial stimuli of the type used in our studies would reflect the more conservative reality. A contrasting perspective, however, is that all attitudes are inherently contextual anyhow and that the accessibility–availability dichotomy is misguided because the notion of stored memories is merely a convenient metaphor, not a literal description (Bohner & Schwartz, 2001; J. P. Mitchell et al., 2003; Searle, 1992; Smith, 1996). If so, then any positive shifts observed would perhaps have to be taken at face value. Moreover, if those attitude shifts were supplemented by correlated changes in overt behavior, then it would be difficult to argue that any changes observed were superficial in the sense of being trivial. Nonetheless, the durability of those attitude shifts would remain an empirical question. It may transpire, for example, that a rich history of primarily positive (or negative) associations toward an object inhibits the extent or longevity of attitude shifts in a negative (or positive) direction brought about by contextual factors. We welcome future experimental research addressing precisely this issue.

[8] We thank an anonymous reviewer for this suggestion.

Conclusion

If automatic attitudes prove to be generally easier to acquire than they are to eliminate, what implications follow? Let us assume that automatic attitudes uniquely predict, and perhaps even prompt, outcomes of consequence, as accumulating research now indicates (Poehlman, Uhlmann, Greenwald, & Banaji, 2004). If such automatic attitudes reflect antisocial prejudices, then the news is bad: people can speedily develop, at an implicit level, unfavorable and undeserved evaluations of social groups that they can only laboriously unburden themselves of them later. The pessimistic prognosis of Devine's (1989) original dissociation model would thereby find some measure of confirmation. On the other hand, if such automatic attitudes reflect prosocial insights, then the news is good: people can speedily develop, at an implicit level, favorable and fitting evaluations of social groups that resist subsequent attempts at uprooting them. The moral of the story would then seem to be that right-minded attitudes should be instilled first before wrong-headed ones gain a foothold, that egalitarian education should begin earlier rather than later so that its beneficial effects can be more far-reaching and enduring. At an implicit level, prevention may be better than cure.

References

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Asendorpf, J. B., Banse, R., & Muecke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behaviour. *Journal of Personality and Social Psychology, 83,* 380–393.

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

Banaji, M. R. (2002). The opposite of a great truth is also true. In J. T. Jost, D. A. Prentice, & M. R. Banaji (Eds.). *The yin and yang of social psychology*: *Essays in honor of William J. McGuire* (pp. 127–140). Washington, DC: American Psychological Association.

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

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48,* 145–160.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology, 62,* 893–912.

Bassili, J. N. (2001). Cognitive indices of social information processing. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 68–88). Oxford, United Kingdom: Blackwell

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin, 91,* 3–26.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7–15.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain, 123,* 2189–2202.

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

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

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81,* 828–841.

Blanton, H., Jaccard, J., & Gonzales, P. M. (2004). *Arbitrary metrics in psychology.* Unpublished manuscript, University of North Carolina, Chapel Hill.

Bohner, G., & Schwartz, N. (2001). Attitudes, persuasion, and behaviour. In A. Tesser & N. Schwartz (Eds.), *Blackwell handbook of social psychology: Individual differences* (pp. 413–435). Oxford, United Kingdom: Blackwell.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind man and the elephant revisited? *Journal of Personality and Social Psychology, 79,* 631–643.

Brauer, M., Wasel, W., & Niedenthal, P. M. (2000). Implicit and explicit components of prejudice. *Review of General Psychology, 4,* 79–101.

Cacioppo, J. T., Marshall-Goodell, B. S., Tassinary, L. G., & Petty, R. E. (1992). Rudimentary determinants of attitudes: Classical conditioning is more effective when prior knowledge about the attitude stimulus is low than high. *Journal of Experimental Social Psychology, 28,* 207–233.

Castelli, L., Zogmaister, C., Smith, E. R., & Arcuri, L. (2004). On the automatic evaluation of social exemplars. *Journal of Personality and Social Psychology, 86,* 373–387.

Chaiken, S., & Trope, Y. (2000). *Dual-process theories in social psychology.* New York: Guilford Press.

Converse, P. E. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In E. R. Tufte (Ed.), *The quantitative analysis of social problems* (pp. 168–189). Reading, MA: Addison Wesley.

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.

Damasio, A. R. (1994). *Descartes" error: Emotion, reason, and the human brain.* New York: Grosset/Putnam.

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

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *Psychology of Evaluation* (pp. 219–244). Mahwah, NJ: Erlbaum.

De Houwer, J., & Eelen, P. (1998). An affective variant of the Simon paradigm. *Cognition & Emotion, 12,* 45–61.

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.

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

Devine, P. G. (2001). Implicit prejudice and stereotyping: How automatic are they? Introduction to the special section. *Journal of Personality and Social Psychology, 81,* 757–759.

Devine, P. G., & Baker, S. M. (1991). Measurement of racial stereotype subtyping. *Personality and Social Psychology Bulletin, 17,* 44–50.

Devine, P. G., Monteith. M. J., Zuwerink, J. R., & Elliott, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60,* 817–830.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82,* 835–848.

Dovidio, J. F., & Gaertner, S. L. (1998). On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. In J. L. Eberhardt & S. Y. Fiske (Eds.), *Confronting racism: The problem and the response* (pp. 3–32). London: Sage.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62–68.

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

Doyle, A. C. (1981). *The complete Sherlock Holmes.* London: Penguin Books.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83,* 1441–1455.

Epstein, S., & Pacini, R. (1999). Some basic issues regarding dual–process theories from the perspective of cognitive-experiential self-theory. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology.* New York: Guilford Press.

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

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87,* 293–311.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior.* Reading, MA: Addison Wesley.

Foroni, F., & Mayr, U. (in press). The power of a story: New, automatic associations from a single reading of a short scenario. *Psychonomic Bulletin and Review.*

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

Gawronski, B., & Strack, F. (2003). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40,* 535–542.

Gawronski, B., Walther, E., & Blank, H. (2004). *The formation of explicit and implicit interpersonal attitudes: On units, sentiments, and cognitive balance.* Unpublished manuscript, Universitat Würzberg.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60,* 509–517.

Glaser, J. (1999). *The relationship between stereotyping and prejudice: Measure of newly formed automatic associations.* Unpublished doctoral dissertation, Yale University.

Glaser, J., & Banaji, M. R. (1999). When fair is foul and foul is fair: Reverse priming in automatic evaluation. *Journal of Personality and Social Psychology, 77,* 669–687.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology, 40,* 357–365.

Greenwald, A. G. (1988). Self-knowledge and self-deception. In J. S. Lockard and D. L. Paulhaus (Eds.) *Self-deception: An adaptive mechanism?* (pp. 113–131). Englewood Cliffs, NJ: Prentice Hall.

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

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science, 283,* 1699–1702.

Greenwald, A. G., Klinger, M. R., & Liu, T. J. (1989). Unconscious processing of dichoptically masked words. *Memory & Cognition, 17,* 35–47.

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Greenwald, A. G., Pickrell, J. E., & Farnham, S. D. (2002). Implicit partisanship: Taking sides for no reason. *Journal of Personality and Social Psychology, 83,* 367–379.

Gregg, A. P. (2000). *The hare and the tortoise: The origins and dynamics of explicit and implicit attitudes.* Doctoral dissertation, Yale University.

Gregg, A. P. (2003). Optimally conceptualizing implicit self-esteem. *Psychological Inquiry, 14,* 35–37.

Gregg, A. P., & Sedikides, C. (2004). *Am I gnasty or gnice? Unmasking the fragility of narcissistic self-regard with the GNAT.* Unpublished manuscript, University of Southampton.

Harmon-Jones, E., & Mills, J. (1999). *Cognitive dissonance progress on a pivotal theory in social psychology.* Washington, DC: Braun-Brumfield.

Hetts, J. J., Sakuma, M., & Pelham, B. W. (1999). Two roads to positive regard: Implicit and explicit self-evaluation and culture. *Journal of Experimental Social Psychology, 35,* 512–559.

Jellison, W. A., McConnell, A. R., & Gabriel, S. (2004). Implicit and explicit measures of sexual orientation attitudes: Ingroup preferences and overt behaviors among gay and straight men. *Personality and Social Psychology Bulletin, 30,* 629–642.

Judd, C. M., & McClelland, G. (1998). Measurement. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 180–232). Boston: McGraw-Hill.

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

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78,* 871–888.

Kim, D. Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66,* 83–96.

Koole, S. L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: Implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology, 80,* 669–685.

Koole, S. L., & Pelham, B. W. (2003). On the nature of implicit self-esteem: The case of the name letter effect. In S. Spencer, S. Fein, M. P. Zanna, & J. M. Olson (Eds.). *Motivated social perception: The Ontario Symposium* (Vol. 9, pp. 93–166). Hillsdale, NJ: Erlbaum.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27,* 313–327.

Kunda, Z., & Thagard, P. (1996).Forming impressions from stereotypes, traits and behaviors: A parallel-constraint-satisfaction theory, *Psychological Review, 103,* 284–308.

Lane, K. A., Mitchell, C. J., & Banaji, M. R. (2004). *Variations in implicit in-group performance: Group membership and group status.* Unpublished manuscript, Harvard University.

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

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic social prejudice. *Journal of Personality and Social Psychology, 81,* 842–855.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the activation of social stereotypes: The moderating role of processing objectives. *Journal of Experimental Social Psychology, 33,* 471–489.

Maison, D., Greenwald, A. G., & Bruin, R. (2001). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin, 32,* 61–69.

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

McConnell, A. R.,& Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

McDell, J. J., Banaji, M. R., & Cooper, J. (2004). *Freedom to choose does not matter to implicit attitudes.* Unpublished manuscript, Harvard University.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26,* 446–456.

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test. *Journal of Experimental Social Psychology, 40,* 366–373.

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation, 23,* 203–217.

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

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

Neumann, R., Hulsenbeck, K., & Seibt, B. (2004). Attitudes towards people with AIDS and avoidance behavior: Automatic and reflective bases of behavior. *Journal of Experimental Social Psychology, 40,* 543–550.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Nosek, B., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition, 19,* 625–666.

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

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math not = me. *Journal of Personality and Social Psychology, 83,* 44–59.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413–417.

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition, 20,* 89–104.

Paulhus, D. L. (1993). Bypassing the will: The automatization of affirmations. In D. M. Wegner and J. W. Pennebaker (Eds.), *Handbook of mental control* (pp. 573–587). Englewood Cliffs, NJ: Prentice Hall.

Perugini, M. (2004). *Individual differences in moral decision-making: Validation of an implicit measure of morality.* Unpublished manuscript, University of Essex, United Kingdom.

Petty, R. E. (1997). The evolution of theory and research in social psychology: From single to multiple effect and process models. In C. McGarty & S. A. Haslam (Eds.), *The message of social psychology: Perspectives on mind in society* (pp. 268–290). Oxford, United Kingdom: Blackwell.

Petty, R. E., & Jarvis, W. B. G. (1998, October). *What happens to the "old" attitude when attitudes change?* Paper presented at the annual meeting of the Society for Experimental Social Psychology, Lexington, KY.

Petty, R. E., & Krosnick, J. A. (1995). *Attitude strength: Antecedents and consequences.* Hillsdale, NJ: Erlbaum.

Phelps, E. A., Cannistraci, C. J., & Cunningham, W. A. (2003). Intact performance on an indirect measure of face bias following amygdala damage. *Neuropsychologia, 41,* 203–208.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12,* 729–738.

Plant, E. A., & Devine, P. A. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75,* 811–832.

Poehlman, T. A., Uhlmann, E., Greenwald, A. G., & Banaji, M. R. (2004). *Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity.* Unpublished manuscript, Yale University.

Richeson, J. A., & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology, 39,* 177–183.

Riketta, M., & Dauenheimer, D. (2003). Manipulating self-esteem with subliminally presented words. *European Journal of Social Psychology, 33,* 679–699.

Rothermund, K., & Wentura, D. (2001). Figure-ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift für Experimentelle Psychologie, 48,* 94–106.

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

Schlenker, B. R. (1975). Self-presentation: Managing the impression of consistency when reality interferes with self-enhancement. *Journal of Personality and Social Psychology, 32,* 1030–1037.

Searle, J. R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT Press.

Seibt, B., Hafner, M., & Neumann, R. (2004). *Prepared to eat: How immediate affective and motivational responses to food cues are influenced by food deprivation.* Unpublished manuscript, Universitat Würzberg.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–447.

Sinclair, L., & Kunda, Z. (1999). Reactions to a Black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77,* 885–904.

Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich & D. Griffin (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379–396). New York: Cambridge University Press.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology, 70,* 893–912.

Smith, E. R., & DeCoster, J. (1999). Associative and rule-based processing: A connectionist interpretation of dual-process models. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 323–336). New York: Guilford.

Spalding, L. R., & Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological Science, 10,* 535–539.

Spencer, S. J., Fein, S., Wolfe, C., Fong, C., & Dunn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24,* 1139–1152.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51,* 165–179.

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology, 50,* 33–48.

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

Teachman, B., & Woody, S. (2003). Automatic processing in spider phobia: Implicit fear associations over the course of treatment. *Journal of Abnormal Psychology, 112,* 100–109.

Teachman, B. A., Gregg, A. P., & Woody, S. (2001). Implicit attitudes toward fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110,* 226–235.

Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology, 73,* 941–959.

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

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude–behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 287–343). Orlando, FL: Academic.

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

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

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.

Appendix

Stimulus lists

Niffite names: Cellanif, Eskannif, Lebbunif, Zallunif

Luupite names: Maasolup, Neenolup, Omeelup, Wenaalup

“Good” words used in the Implicit Attitudes Test (IAT): *excellent, heaven, joy, trust, peace, enjoyment, friend, honest, sweetheart, love, freedom, paradise*.

“Bad” words used in the IAT: *murder, cancer, war, disaster, hatred, slaughter, bomb, agony, torture, slime, filth, traitor*.

Positive primes used the supraliminal priming procedure**:** *good, honest, honorable, benevolent, peaceful, principled, cultured, law-abiding, respectable, trustworthy, likeable, friendly*.

Negative primes used in the supraliminal priming procedure**:** *treacherous, evil, vicious, sadistic, bloodthirsty, murderous, savage, barbaric, vindictive, depraved, sickening, malicious*.

Received November 7, 2002
Revision received June 16, 2005
Accepted June 20, 2005 ■