Draft of January 6, 2001

A UNIFIED THEORY OF IMPLICIT ATTITUDES, STEREOTYPES, SELF-ESTEEM, AND SELF-CONCEPT

Anthony G. Greenwald, University of Washington

Mahzarin R. Banaji, Yale University

Laurie A. Rudman, Rutgers University

Shelly D. Farnham, University of Washington

Brian A. Nosek, Yale University

Deborah S. Mellott, University of Washington

Acknowledgment. This research was supported by grants from National Science Foundation, SBR-9422242, SBR-9710172, SBR-9422241, and SBR-9709924 and from National Institute of Mental Health, MH-41328, MH-01533, and MH-57672. The authors thank Eugene Aidman, John Bargh, Richard Gonzalez, Mary Lee Hummert, Chester Insko, John Kihlstrom, Eliot Smith, Mark VandeKamp, Vivian Zayas, and three anonymous reviewers for comments on earlier drafts. Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, WA, USA, 98195-1525. Electronic mail may be sent to agg@u.washington.edu.

Abstract. This theoretical integration of social psychology's main cognitive and affective constructs was shaped by three influences: (a) recent widespread interest in automatic and implicit cognition, (b) development of the Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998), and (c) social psychology's consistency theories of the 1950s – especially Heider's (1958) balance theory. The *balanced identity design* is introduced as a method to test correlational predictions of the theory. Data obtained with this method revealed that predicted consistency patterns were strongly apparent in the data for implicit (IAT) measures, but not in those for parallel explicit (self-report) measures. Two additional not-yet-tested predictions of the theory are described.

The Cognitive Consistency Theoretical Tradition

Theories of cognitive consistency dominated social psychology in the 1960s. The most influential ones had appeared in the 1950s, including Osgood and Tannenbaum's (1955) congruity theory, Festinger's (1957) cognitive dissonance theory, and Heider's (1958) balance theory. The high point of consistency theory was the 1968 publication of the 6-editor, 920-page handbook, *Theories of Cognitive Consistency: A Sourcebook* (Abelson, Aronson, McGuire, Newcomb, Rosenberg, & Tannenbaum, 1968); it contained 84 chapters by 75 contributing authors. Now, just over thirty years later, it is remarkable that these once-dominant theories receive at most occasional mention by social psychologists. There are several ways to understand this fall from favor.

1. *Ascent to common sense wisdom*. In part, reduced attention to consistency theories may be due to their having been so thoroughly woven into the fabric of social psychology as to have acquired the character of unquestioned wisdom, no longer needing research investigation.

2. *Unresolved competition among theories*. Competition among consistency theories in the 1960s and 1970s drew attention more to their peripheral theoretical differences than to their central areas of agreement. Research on these disagreements never produced a decisive preference among the theories. This lack of resolution could have created the impression that all of the theories had problems. At the least, it could have diverted attention from their common theoretical core, which was never contested.

3. *Rise of attribution theory*. Festinger's cognitive dissonance theory drew attention to the dramatic problem of understanding cognitive change following 'forced compliance' (Festinger & Carlsmith, 1959). Even though dissonance theory was built on a seemingly rational foundation of cognitive consistency, its interpretation of forced compliance was counterintuitive and puzzling — it confronted then-dominant learning theories by predicting that smaller incentives would produce greater liking. This confrontation focused enormous research attention on the forced-compliance problem in the 1960s and 1970s, partly in its incarnations as *induced compliance* and *counterattitudinal role playing*. This research activity in turn shifted the center of theoretical action from the competition among cognitive consistency theories to a competition among dissonance theory and numerous rivals to explain the induced-compliance results. The set of competitors included Bem's (1972) self-perception theory, Jones and Davis's (1965) correspondent inference theory, Tedeschi, Schlenker, and Bonoma's (1971) impression management theory, and especially Kelley's (1967) attribution theory, which was itself profoundly shaped by Heider's work. Attribution theory became the dominant approach of the 1970s.

4. *Limited success of application attempts*. Remarkably, the set of consistency theories generated few practical applications. Of the three major consistency theories, cognitive dissonance theory might have been thought to offer the greatest promise of interesting and unexpected applications, because of the often nonobvious character of its predictions. However, even after more than 40 years of research, adherents of cognitive dissonance theory continue to debate about how best to state the theory and about how to translate it into effective applications (Harmon-Jones & Mills, 1999).

5. *Reliance on self-report measures*. The era of dominance of cognitive consistency theories was also a period during which social psychological research method depended almost exclusively on self-report measures of cognitive and affective constructs. There are well known problems with self-report measures, because they depend crucially on (a) subjects' willingness to report private knowledge, and (b) subjects' ability to report such knowledge accurately. Consequently, self-report measures can go astray when respondents are either unwilling or unable to report accurately. These problems could be more than enough to obscure the operation of consistency processes.

The foregoing five points notwithstanding, consistency theories have left a strong imprint on social psychology. The deepest impression may be on the way in which psychologists now understand how incentives (rewards and punishments) function in human behavior. Prior to 1960, psychology was dominated by noncognitive law-of-effect conceptions of reward and punishment, based on the learning-theory traditions of (among others) Thorndike, Hull, and Skinner. The consistency theories, and especially cognitive dissonance theory, replaced this view with thoroughly cognitive conceptions of reward and punishment effects.

Indirect Measurement Strategies

At about the same time that consistency theories rose to prominence in the 1960s, self-report measures were being heavily attacked because of their susceptibility to such artifacts as demand characteristics (Orne, 1959), evaluation apprehension (Rosenberg, 1969), and impression management (Tedeschi, Schlenker, & Bonoma, 1971; Weber & Cook, 1972). Complementing these empirical attacks on self-report measures, Nisbett and Wilson (1977) offered a theoretical and methodological critique of the flawed introspectionism of self-report methods.

During the period of concerted critique of self-report measures in the 1960s and 1970s, social psychologists were attracted to indirect measures, sometimes referred to as nonreactive or unobtrusive measures (Webb, Campbell, Schwartz, & Sechrest, 1966; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Interestingly, social psychology's attraction to indirect measures in the 1970s and 1980s proved to be little more than flirtation. Perhaps the appeal of these measures was undermined both by their labor-intensive character and by their seeming remoteness from the cognitive constructs to which social psychologists were increasingly drawn in the late 20th century.

In the late 1980s and early 1990s, useful and efficient alternatives to self-report measures began to appear in social cognition research. Several of the new indirect measures were inspired by developments in implicit cognition research (e.g., Jacoby, Lindsay, & Toth, 1992; Schacter, 1987). A significant justification for these measures was the widely shared belief that they provided access to a cognitive domain that was not reached by self-report measures (Bargh, 1997; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald & Banaji, 1995). Subsequently, Greenwald, McGhee, and Schwartz (1998) introduced the Implicit Association Test (IAT), which provided the measures of implicit social cognition constructs that were used to test the theory developed in this article.

A Goal of Theoretical Unification

The present theory started, not from an interest in consistency theories, but as an attempt to understand results obtained in the first few years of research using the IAT method. One especially intriguing result was an unexpected finding that women implicitly associated *female* with *strength* to about the same extent that they associated *male* with *strength* (Rudman, Greenwald, & McGhee, in press, Experiment 1). It had been expected, instead, that women would implicitly associate male (more than female) with strength — that is, that women would show the same gender-stereotypic association of *male* with *strength* that men did. In the experiment that first observed this sex difference in the gender-potency stereotype (Rudman et al., in press, Experiment 1), the words that represented *strong* in the IAT were evaluatively more positive than the words representing *weak*. A possible interpretation, therefore, was that women's implicit female-strong association might reflect their self-esteem. That is, female-strong might have been the cognitively consistent product of self-female, self-positive, and strong-positive associations – none of which was measured in the experiment. Followup studies that used multiple IAT measures eventually crystallized into the *balanced-identity* design, which is described and illustrated in this article. The theory statement in this article thus developed partly as a post hoc interpretation of findings obtained with the balanced identity design, but it extends further to include principles that have not yet been subject to experimental test.

Primitive Terms

The present theory uses three terms that are left without precise definition – *concept*, *association strength*, and *concept activation*. Because these terms are familiar from a long history of use in psychology, their informal use here should not create problems. The following paragraphs explain the loose meanings attached to these three terms in this article.

Concept. Concepts that are significant for this theory represent persons, groups, or attributes. Among attribute concepts, positive and negative valence are especially important. Concepts may be reduced to further primitives — for example, to structured relations among prototypes or exemplars (Smith & Zarate, 1992) or, at a lower level, among the features that compose concept instances (e.g., Smith & Medin, 1981). Such added details are not specified here out of a conviction that they do not have substantial consequence for the assumptions and principles that follow.

Association strength. Associations are relations between pairs of concepts that can be represented by familiar node (= concept) and link (= association) diagrams (e.g., present Figure 1). Strength of association is understood as the potential for one concept to activate another (see just below). The theory in this article treats associations as bidirectional, facilitatory, and continuously variable in strength. The theory would not likely be changed in any essential way if it were to use alternative conventions for association strength (e.g., with associations being unidirectional and/or both facilitatory and inhibitory and/or all-or-none in strength).[1]

Concept activation. Concepts are assumed to be activated either by external stimuli or by excitation via their associations with other, already active, concepts. The assumption that associations are strengthened between two simultaneously active concepts was prominent in Hebb's (1949) theorizing and continues to be relied on in modern neural network (connectionist) theories, while also providing a basis for the association-strengthening postulated in the present theory's first principle.

Definitions

The theory defines four familiar social-cognitive constructs in terms of associations among concepts.

An **attitude** is the association of a social object or social group concept with a valence attribute concept.

A **stereotype** is the association of a social group concept with one or more (non-valence) attribute concepts.

Self-esteem is the association of the concept of self with a valence attribute.

---

[1]This assertion is made in the expectation that process assumptions can be used flexibly enough to allow different representation formats to function equivalently (as described, e.g., by Anderson, 1978). A reviewer suggested that the facilitatory-association-only assumption would have difficulty in describing the relation between opposites such as *hot* and *cold*, or between related, but affectively incompatible, pairs such as *Hitler* and *Jews*. Such cases can, however, be accommodated in a facilitatory-only association scheme with the aid of a process assumption such as the present theory's *imbalance-dissonance* principle (see below), which describes resistance to establishment of direct associations between members of such pairs.

A **self-concept** is the association of the concept of self with one or more (non-valence) attribute concepts.

Defining each of these four constructs in terms of associations makes it possible to describe relations among the four constructs with a small set of theoretical principles.

Assumptions

Like the three primitive terms, the following three assumptions stray little from views that are widely shared among social psychologists. They are, in effect, pieces of the paradigmatic common ground of modern social psychology.

Associative social knowledge. We assume that an important portion of social knowledge can be represented as a network of variable-strength associations among person concepts (including self and groups) and attributes (including valence).

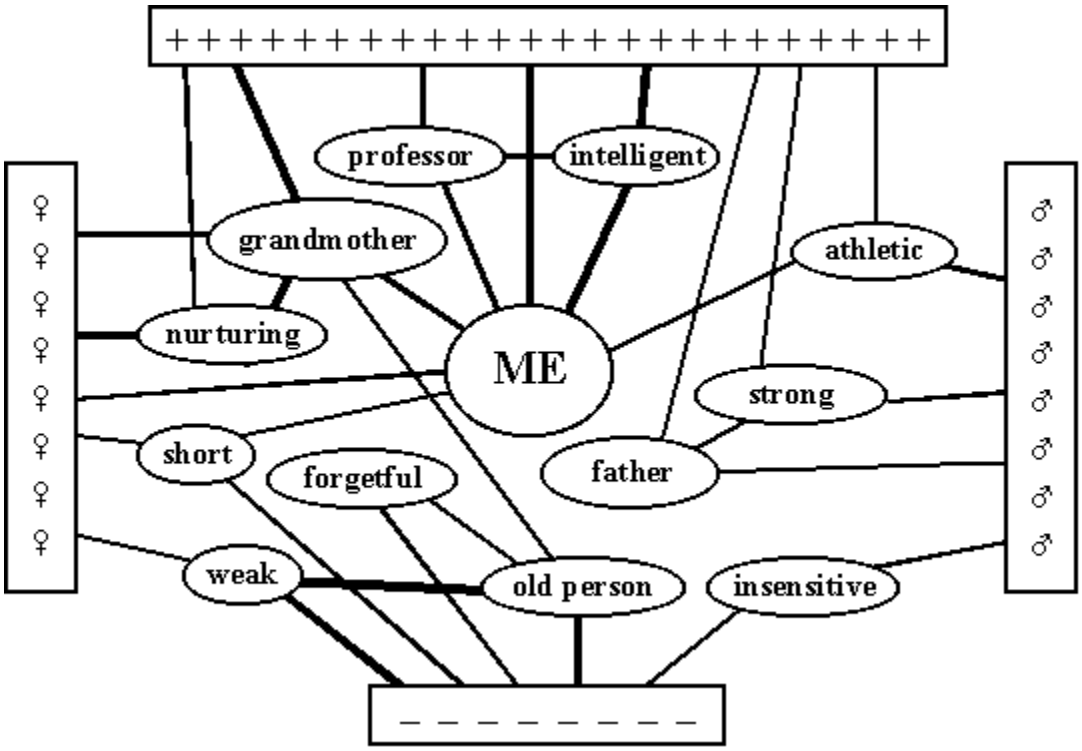

Figure 1. **A Social Knowledge Structure (SKS)**. This structure includes associations that correspond to social psychological constructs of self-concept, self-esteem, stereotype, and attitude in the psyche of an elderly female academic. Nodes (ovals) represent concepts, and links (lines) represent associations. Line thickness represents strength of association. The **self-concept** includes links of the *Me* node to concepts that include roles (professor, grandmother) and trait attributes (intelligent, athletic); **self-esteem** is the collection of associations — either direct or mediated via components of the self-concept — of the *Me* node to valence (+ + + or – – –); **stereotypes** are associations of group concepts such as old person, grandmother, professor, male (♂♂♂), and female (♀♀♀) with attribute concepts; and **attitude** is the collection of links, either direct or mediated via components of a stereotype, that connect a social concept to valence.

Centrality of self. Following Koffka (1935), much recent work has identified *self* as a central entity in the structure of social knowledge (e.g., Greenwald, 1981; Greenwald & Pratkanis, 1984; Kihlstrom & Cantor, 1984; Kihlstrom & Klein, 1994). In an associative knowledge structure, self's centrality can be represented by its being associated with many other concepts that are themselves highly connected in the structure.

Self-positivity. Because valence is represented as an attribute concept in the associative structure of social knowledge, self-esteem can be represented as a connection of the self node to a valence node. The frequent empirical observation that self-esteem is positive in normal populations translates to an assumption that, for most people, the self node is associated with the positive valence node.

Three Definitions and Three Principles

Figure 1 displays a schematic social knowledge structure (SKS) that incorporates the theory's primitive terms and assumptions. Although Figure 1 includes only a tiny fraction of the concepts (objects and attributes) of any actual social knowledge structure, it nevertheless includes structures corresponding to the theoretical constructs of *self-concept*, *self-esteem*, *stereotype*, and *attitude* (see Figure 1's caption).

The present theory will gain the ability to describe relations among self-esteem, self-concept, stereotypes, and attitudes by stating three principles that constrain associative strengths within structures such as Figure 1's SKS. Each of the theory's three principles is named in a way that identifies a debt to the tradition of cognitive consistency theories, and each uses a preliminary definition to describe a theoretically relevant property of the knowledge structure.

Definition 1. Shared first-order link. When each of two nodes is linked to the same third node, the two are said to have a shared first-order link.

Principle 1. Balance-congruity. When two unlinked or weakly linked nodes share a first-order link, the association between these two should strengthen.

Principle 1 was named to acknowledge its debts to both Heider's (1946, 1958) balance theory and Osgood and Tannenbaum's (1955) congruity theory. In Figure 1's structure, the balance-congruity principle should tend to strengthen (among others) several links involving the *Me* node, including *Me-father*, *Me-male*, *Me-nurturing*, *Me-old person*, and *Me-weak*. Importantly, all but one (*Me-nurturing*) of these possible new links is opposed by the next principle.

Definition 2. Bipolar opposition of nodes. To the extent that two nodes have fewer shared first-order links than expected by chance, they can be described as bipolar-opposed.

As diagramed in Figure 1, SKS has two prominent pairs of bipolar-opposed nodes, those for valence (positive, negative) and sex/gender (male, female). (SKS contains one other bipolar pair — weak and strong — and could easily be extended to include others, such as intelligent-stupid, short-tall.)

Principle 2. Imbalance-dissonance. The network resists forming new links that would result in a node having first-order links to both of two bipolar-opposed nodes.[2]

Principle 2 is named to acknowledge its debt to both Heider's (1958) balance theory and Festinger's (1957) dissonance theory. The resistance to new links embodied in the imbalance-dissonance principle is theoretically necessary to oppose the otherwise inevitable effect of the balance-congruity principle, in conjunction with environmental influences, to produce links among all pairs of nodes. In Figure 1, for example, *Me* is linked to *female* and to *athletic*, the latter of which is linked to *male*. The imbalance-dissonance principle resists the formation of a link of *Me* to *male*, which would otherwise be called for by the balance-congruity principle, operating on the shared first-order link of both *Me* and *male* to *athletic*. Similarly, in Figure 1, the imbalance-dissonance principle resists influences that might strengthen *Me-negative* (e.g., via the shared first-order link to *short*).

The imbalance-dissonance principle functions to avoid configurations that link any node to both of two bipolar-opposed nodes. An additional principle with similar function is needed for situations that involve sustained external pressure toward generating an imbalanced configuration. As example, consider a situation in which one's loved sibling (*A*) gets married to person *B*, who happens to be a criminal (*C*). The existing association of *A* to positive valence should produce (via the assumed new link of *A* to *B*, in conjunction with the balance-congruity principle) a link of *B* to positive valence. At the same time, the likely unalterable association of the criminal concept (*C*) with negative valence should (again, by virtue of balance-congruity) tend to produce an association of *B* to negative valence. The resulting tendency for *B* to develop links to bipolar-opposed nodes (positive and negative valence) is opposed by the imbalance-dissonance principle. In this situation, a structural adaptation that can avoid the sustained confrontation of imbalancing influences would be useful. The third principle provides this.

Definition 3. Pressured concept. A concept is pressured when sustained or repeated influences should cause it (via the balance-congruity principle) to develop links to both of two bipolar-opposed nodes.

Principle 3. Differentiation. Pressured concepts tend to split into subconcepts, each linked to a different one of the pressuring bipolar-opposed nodes.

---

[2]For this article, the concept of bipolar opposition does not need more precise statement than Definition 2. Nevertheless, to indicate how greater precision might be achieved, consider that the opposition of two nodes can be quantified in terms of their number of shared first-order links relative to the total number of links in which they participate. For example, consider the *weak* and *nurturing* nodes in SKS (Figure 1). These two nodes have four shared first-order links relative to eight total links (four for *weak* and four for *nurturing*). Sharing four of eight first-order links is substantial, given that the expected number of shared first-order links for *weak* and *nurturing* is just over one. (To calculate: there are 15 other nodes with which these two might have shared first-order links. With each of *weak* and *nurturing* having four direct links, the expected number of shared first-order links is therefore $4/15 \bullet 4/15 \bullet 15 = 1.07$.) By contrast, *strong* and *weak* have zero shared first-order links relative to six total links (four for *weak* and two for *strong*). Although measures of bipolar opposition should be based on a more complete specification of associative structure than provided by the deliberately simplified structure of Figure 1, the patterns just described indicate that *weak* and *nurturing* are positively associated (even though there is no direct link between them), whereas *weak* and *strong* are possibly in bipolar opposition. Returning to the point made in Footnote 1, opposites such as *hot* and *cold* would share at least one first order link (e.g., to their common superordinate, the abstract concept of *temperature*) but likely would not have as many shared first-order links as would synonyms, *hot* and *warm*, and – due to theorized operation of the imbalance-dissonance principle.

Principle 3's name came from Heider's (1958) analysis of a similar situation (see Figure 2). In the example of the sibling's criminal spouse, the spouse (*B*) is a pressured concept. This pressure would be removed if Person *B* could split into two concepts, one linked to negative valence (e.g., *B*'s past criminal identity) and the other to positive valence (*B*'s current identity as loving spouse). This split, or differentiation, removes pressures toward change that result from the balance-congruity principle and are resisted by the imbalance-dissonance principle. The differentiation principle embodies the cognitive operation that is known as *subtyping* in research on stereotypes (e.g., Deaux, Winton, Crowley, & Lewis, 1985; Hewstone, Macrae, Griffiths, & Milne, 1994; Weber & Crocker, 1983).

Similarities to Heider's Balance Theory

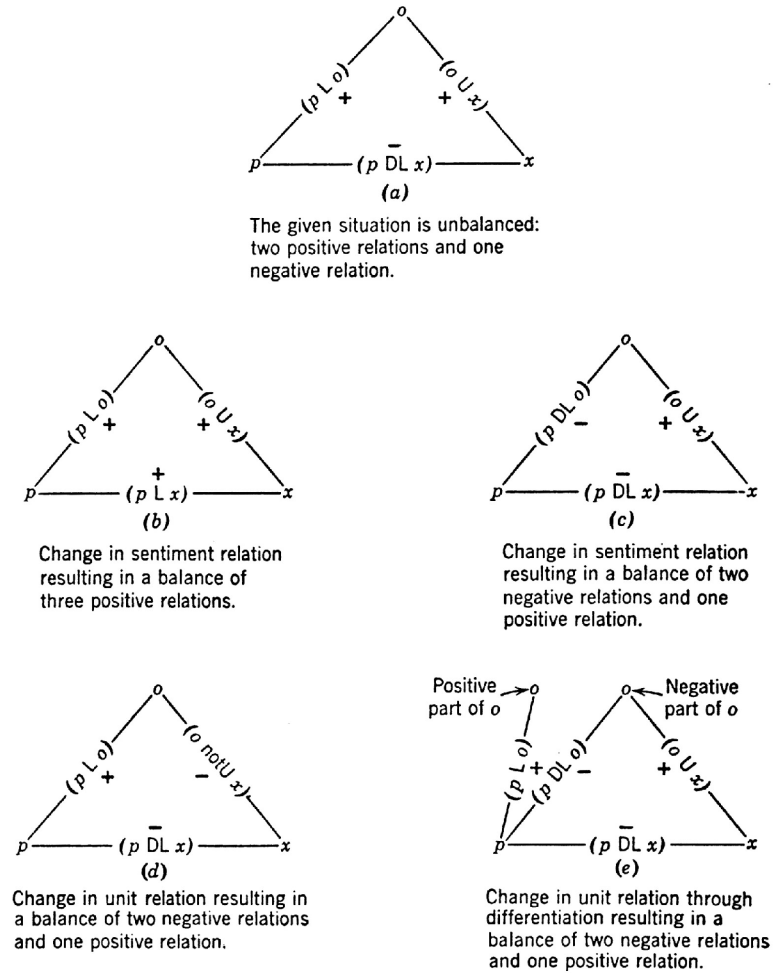

Figure 2. **Heider's representation of consistency principles**. This figure reproduces Heider's portrayal of imbalance (*a*) and balance (*b*–*e*) that appeared in Figure 1 on p. 208 of his chapter on 'Sentiment' in *The Psychology of Interpersonal Relations* (Heider, 1958). *p* = person; *o* = other; *x* = concept; L = positive sentiment relation; DL = negative sentiment relation; U = unit relation.

There are several similarities of the present theory to Heider's balance theory. Insights corresponding to all of Principles 1-3 were captured in Heider's (1958) diagrams of balanced and imbalanced configurations for sentiment and unit relations. In Heider's diagrams (reproduced here as Figure 2) the balance-congruity principle appears in the balanced structures *b–d*, the imbalance-dissonance principle in diagram *a*, and the differentiation principle in diagram *e*. The chief differences between Heider's representations and those of the present theory are that (a) Heider limited attention to links that involved a person object (either self [*p*] or other [*o*]), (b) Heider distinguished *unit* (association) from *sentiment* (liking) links, in contrast to the present theory's use of just one (associative) type of link, (c) Heider focused more on the role of consistency in modifying existing links than on its role in creating (or avoiding) new links, and (d) Heider did not distinguish between unassociated and bipolar-opposed pairs of nodes.[3]

In order to represent the complexity of consciously construed relations among psychological objects, Heider focused on person-object relations and distinguished unit from sentiment relations. Heider's observation that many person-object relations could be described using just the unit and sentiment relations was a remarkable and theoretically effective simplification. The present theory uses an even more radical simplification to obtain even broader scope — collapsing both (a) the distinction between person concepts and other concepts and (b) the distinction between unit and sentiment relations. This step has been influenced by modern connectionist and neural network modeling, themselves forms of theory that reduce mental representations to node and link structures (Rumelhart & McClelland, 1986; Smith, 1996).

Methods for Empirical Tests

The Implicit Association Test (IAT):

Measuring Associative Strength in the Social Knowledge Structure

The present theory's predictions can be tested in studies that use self-report measures of the types widely used in social psychology of the last several decades. However, for two reasons, self-report measures are not necessarily preferred for tests of the present theory. First, some of the associative links of SKS may not be available to introspection and may therefore not permit accurate assessment by self-report measures (cf. Greenwald & Banaji, 1995). Second, self-report measures are susceptible to artifacts (such as impression management and demand characteristics) that can distort reporting even of associations that are introspectively available. Consequently, in the experiments reported here the unified theory's predictions have been tested not only with self-report measures, but also with a recently developed indirect measurement method, the Implicit Association Test.

Format of the measure. The Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998) indirectly measures strengths of associations between concepts. In the IAT's procedure, subjects are asked to sort stimuli representing four concepts into just two response categories, each of which includes two of the four concepts. Usefulness of the IAT to measure association strength depends on the assumption that when the two concepts that share a response are strongly associated, the sorting task is considerably easier than when the two response-sharing concepts are either weakly

[3]Heider's struggle with the difficulty of not having a distinction like that between unassociated and bipolar-opposed nodes can be seen in his discussion of 'some difficulties connected with the notU [i.e., not-unit] relation' (Heider, 1958, pp. 201-202).

associated or bipolar-opposed. The IAT illustrated in Figure 3 uses the four concepts of male, female, self, and other to provide a self-concept measure of gender identity as male or female. Subjects who identify as female should find the IAT's task easier when the two sorting categories are *female*-or-*self* vs. *male*-or-*other* than when the two sorting categories are *male*-or-*self* vs. *female*-or-*other*.[4]

Concept Contrasts for the Gender Self-Concept IAT

	Male versus Female		Self versus Other	
concepts	Male	Female	Self	Other
items	male man boy he sir	female woman girl she lady	I me my mine self	they them their theirs others

Sequence of Tasks for Gender Self-Concept IAT

Steps	Concepts for left response	Concepts for right response
1	Male	Female
2	Self	Other
3	Self or Male	Other or Female
4	Female	Male
5	Self or Female	Other or Male

Figure 3. **Illustration of the Implicit Association Test to measure gender self-concept**. The IAT starts by introducing subjects to the four concepts that will be used in a series of 5 tasks. In this illustration, one pair of concepts is introduced in the first task by asking subjects to respond with left key to words representing *male* and with right key to words representing *female*. In the second task, the second pair of concepts is introduced, with subjects asked to respond left to words representing *self* and right to words representing *other*. The third step introduces a combined task, in which words representing either *male* or *self* get the left response and words representing either *female* or *other* get the right response. The fourth task reverses the first, and the fifth task combines the tasks of the 2nd and 4th steps. The IAT effect measure is constructed by comparing performance in the 3rd and 5th steps. If the subject responds more rapidly in the *male*-or-*self* vs. *female*-or-*other* task than in the *female*-or-*self* vs. *male*-or-*other* task, this indicates that, in combination, the *male-self* and *female-other* associations are stronger than the *female-self* and *male-other* associations.

In its typical uses, the IAT measures *relative* strengths of associations in a structure such as SKS. In the Figure 3 example, rather than providing an absolute measure of strength of any individual associative link, the IAT provides a measure (the 'IAT effect') that compares the combined strength

[4]The notation '*female*-or-*self*' indicates that items representing the concepts *female* and *self* are sorted together (the subjects gives the same keyboard response to any item representing either concept) in the IAT.

of *female-self* and *male-other* to the combined strength of *male-self* and *female-other*. This relative-strength indicator is especially useful for testing the present theory's predictions that compare the strengths of associations of valence or self with bipolar-opposed pairs of concepts.

Properties of the IAT measure. Among the first goals of research using the IAT were to establish that the measure could detect valence differences that were either (a) almost universal in the population (e.g., preference for flowers over insects) or (b) expected as differences between subject populations (e.g., between Korean Americans and Japanese Americans in valences associated with their respective ethnicities). These demonstrations were provided by Greenwald et al. (1998), who additionally demonstrated that the IAT was free of several possible sources of procedural artifact. In particular, the IAT effect measure was uninfluenced by whether the pleasant category was assigned to left hand or right hand or by variations (ranging from 150 ms to 750 ms) in the interval between successive trials. Further, effects obtained with the IAT were quite robust over variations in the manner of treating data from incorrect responses and from non-normal response latency distributions. Subsequent research has extended evidence for the IAT's internal validity by establishing that the IAT's association measures are not influenced by variation in familiarity of items used to represent contrasted attitude-object concepts (Dasgupta, McGhee, Greenwald, & Banaji, in press; Ottaway, Hayden, & Oakes, in press; Rudman, Greenwald, Mellott, & McGhee, in press).[5]

Greenwald et al. (1998) reported an influence due to the order of administering the two critical IAT tasks (i.e., Tasks 3 and 5 in Figure 3). Performance on either task tends to be faster when it is performed third in order, rather than fifth in order in Figure 3's sequence. This procedural effect has been accommodated in subsequent research chiefly by counterbalancing the two possible orders of these two tasks.

Several studies have demonstrated sensitivity of IAT measures to experimental manipulations that might be expected to influence automatic expressions of attitudes and stereotypes. Dasgupta and Greenwald (in press) demonstrated that exposure to admirable exemplars of stigmatized categories (African American and elderly) reduced implicit negativity toward those categories. Haines (1999) found that women's being assigned to a powerful role in a simulation game increased the IAT-measured association of self with strength. Blair, Ma, & Lenton (in press) showed that a guided exercise of imagining a strong woman decreased an IAT measure of the gender stereotype that associates male, more than female, with strength. Rudman, Ashmore, & Gary (1999) showed reduced IAT-measured implicit prejudice in students enrolled in a Prejudice and Conflict seminar. Karpinski and Hilton (in press) demonstrated that presenting 200 word pairs that linked the word 'elderly' to various pleasant words and the word 'youth' to various unpleasant words reduced the magnitude of an otherwise strong IAT effect that indicated automatic preference for young over old.

In order to be used in testing the present theory's predictions, the IAT must be sensitive to individual differences, over and above its demonstrated sensitivity to group differences. Test-retest reliabilities of IAT measures, observed in as-yet-unpublished studies, have averaged approximately $r = .6$ (e.g., Bosson, Swann, & Pennebaker, 2000; Dasgupta & Greenwald, in press; Greenwald & Farnham, 2000). Theoretically interpretable within-group individual differences have been observed

[5] A limit on this generalization about the IAT's immunity to familiarity effects occurs when the IAT includes an artificial concept that is composed totally of unfamiliar and meaningless items, such as nonsense words. In tests involving associations with valence, such pseudo-concepts produce data indicating that they have negative valence. This may be accurate, but it may be more appropriate to suggest that the IAT should not be used to assess strengths associations that involve such vacuous concepts.

by Greenwald et al. (1998), Rudman, Ashmore, and Gary (1999), Rudman and Glick (in press), and Rudman, Greenwald, & McGhee (in press). For example, prejudice against female job applicants is associated with IAT-assessed (but not explicit) gender stereotypes (Rudman & Glick, in press). Correlations of IAT measures with semantic priming measures of association strengths show that these two procedures converge as measures of strength of automatic associations (Cunningham, Preacher, & Banaji, in press; Mellott, Cunningham, Rudman, Banaji, & Greenwald, in preparation; Rudman & Kilianski, 2000). Also, IAT-assessed implicit prejudice has been shown to correlate with fMRI-assessed activation of the amygdala (a subcortical structure associated with emotional learning and evaluation) in White subjects exposed to unfamiliar Black faces (Phelps et al., 2000).

The Balanced Identity Design

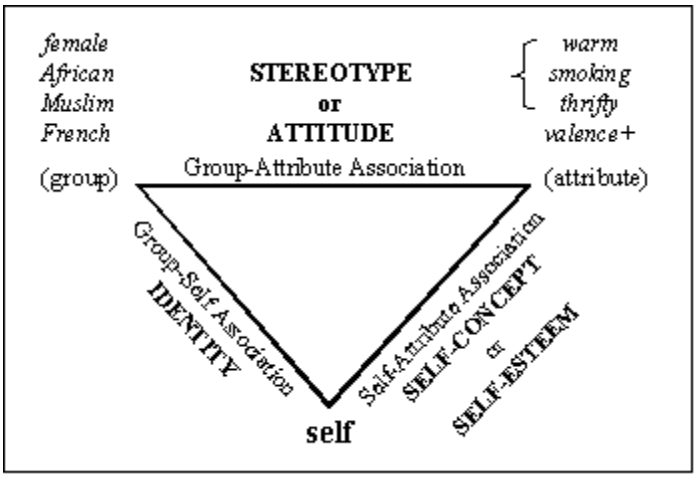

Figure 4. **A representation format for balanced identity designs**. Each vertex of the triangle represents a *concept*. A balanced identity design always includes *self* as one of the concepts (bottom vertex), and also includes both a social category (group) concept and an attribute concept. In italics, above the group and attribute vertices, are examples of concepts that could play those roles in the design. The three associations measured in the design are identified on the triangle edges that join the vertices for the two associated concepts. The group-self association corresponds to an *identity*. The labels for the other two types of associations depend on whether the attribute is valence or not. If the attribute is valence, then the group-attribute association is an *attitude* and the self-attribute association is *self-esteem*. If the attribute is not valence (e.g., any of the bracketed three at upper right), then the group-attribute association is a *stereotype* and the self-attribute association is an aspect of *self-concept*.

The present theory defines four social-cognitive constructs — self-esteem, self-concept, stereotypes, and attitudes — as associations among concepts. The theory's three principles generate predicted relationships among measures of these associations. Research aimed at testing the theory introduces a class of *balanced identity* research designs that are identified by four features: (a) Examination of a triad of potentially associated concepts – always including *self* – and also including a *social category* and an *attribute*, (b) measurement of the three associations that link all pairs of concepts in this triad, (c) obtaining data from subjects who are expected to vary in strength of the association between self and the social category (i.e., varying strengths of *identity*), and (d) use of

statistical tests for predicted patterns that involve the three associations simultaneously. 'Identity' is in the name of the design because the design always includes an identity association, which is the association of self with a social category. 'Balanced' is in the name of the design because these designs can reveal consistency, of the sort hypothesized in Heider's balance theory, within the triad of associations. Figure 4 introduces a representational format for balanced identity designs. The class of balanced identity designs is potentially enormous because of the many possible ways of selecting social categories and attributes for investigation.

Balanced identity designs are well suited to testing predictions derived from the present theory's balance-congruity principle, which holds that two concepts with a shared first-order link should develop a mutual association. Thus, for *self* and *positive* valence, along with any *group* concept, the existence of both *self-positive* and *self-group* associations constitutes a network fragment with a shared first-order link — both *group* and *positive* are linked to *self*. For this configuration, the balance-congruity principle predicts that the attitude toward the group (i.e., the *group-positive* association) should develop in proportion to the product of the strengths of the *self-positive* and *self-group* links. In other words, groups associated with self should share in self's valence.

Figure 5 illustrates an associative structure corresponding to a possible balanced identity design in the domain of women's gender self-concept. As shown in the figure, if *self* ('Me') has the expected (for women) associations with both *positive* (valence) and *female*, the conditions exist for development of a balancing association of *positive* with *female*, by operation of the balance-congruity principle. Prediction 1 states the result of this derivation so as to encompass any configuration involving self, positive valence, and a membership group (i.e., an ingroup).

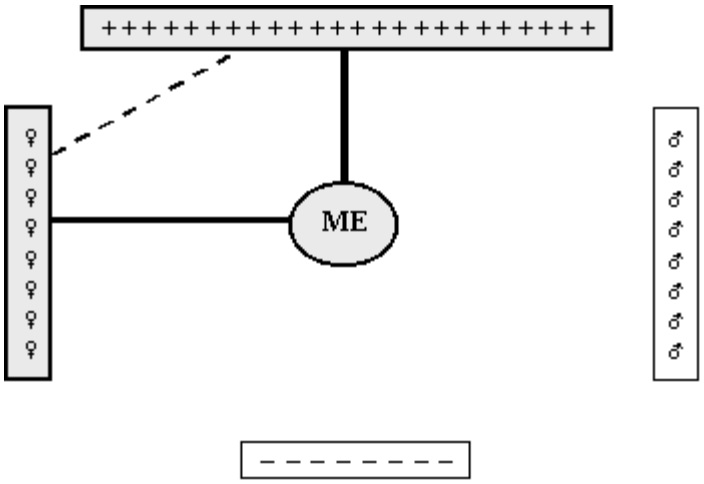

Figure 5. **Balance of identity and attitude**. The diagram shows a fragment of a woman's social knowledge structure in which *positive* and *female* have a shared first-order link to self (*Me*). The *Me-positive* link represents self-esteem, and the *Me-female* link represents ingroup identity as female. In this situation, the balance-congruity principle (strengthening of association between two nodes that have a shared first-order link) calls for strengthening the *female-positive* link, indicated by the dashed line.

Prediction 1: **Balanced identity and attitude**. *Ingroup attitude (ingroup-positive association) should be a multiplicative function of the strengths of ingroup identity (self-ingroup association) and self-esteem (self-positive association)*

The multiplicative form of Prediction 1 follows from the balance congruity principle's appeal to the notion of a *shared* first-order link. That is, if either the *self-positive* or the *self-ingroup* link is of zero strength then there is (a) no shared link and (b) no tendency to form the third (*ingroup-positive*) link.

Statistical Analysis of the Balanced Identity Design

Figure 6 describes a theoretically expected data pattern for any set of three variables that have the interrelationships described in Prediction 1 — one variable (Criterion) being a multiplicative product of the other two (Predictors A and B). Prediction 1 could have been stated with strength of any of the three associations it mentions expected to be a multiplicative function of strengths of the other two. Accordingly, any of the three association measures in a balanced identity design can be in the role of Figure 6's Criterion. The three variables are thus effectively interchangeable in their roles in data analysis.

In the experiments presented in this article, the three associations of each balanced identity design are measured on numeric scales. In order to permit graphic presentation, one of the three variables in Figure 6 (arbitrarily, Predictor B) is treated as an index variable with three levels — low, moderate, and high. Figure 6 displays the expected regression of the Criterion variable on Predictor A separately for these three levels of Predictor B. Because the variables of a balanced identity design can take any of the three roles in Figure 6, the implication of Figure 6 is that the slope of the regression relation between any two variables (e.g., Criterion and Predictor A) is governed by the level of the third variable (Predictor B). When the third variable is at a high level, the expected relationship between the first two variables is positively sloped; when the third variable is at a low level, the expected relationship between the first two is negative.

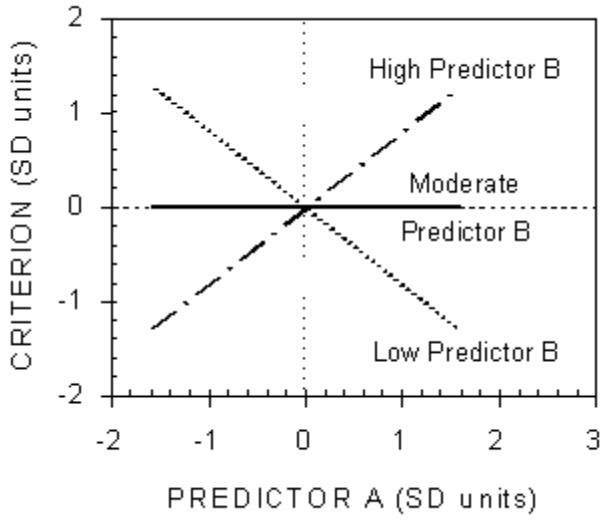

Figure 6. **Expected data pattern for balanced identity designs**. When the three measures of the balanced identity design vary through their full ranges, Prediction 1 calls for the finding of an interaction effect in the regression of any one variable on the other two. The interaction effect is represented here by three different slopes for the regression of a Criterion on one predictor (A) for low, moderate, and high levels of a second predictor (B).

Multiple regression analysis. The pattern shown in Figure 6 has a strong implication for the results of a multiple regression analysis in which any of the three variables of the balanced identity design is criterion, and the other two are predictors. In particular, the data should be fit entirely by the interaction effect in the first step of a 2-step hierarchical analysis that (a) includes only the interaction effect term in the first step and (b) adds the interaction's two component variables as separate predictors in the second step. Further, the regression coefficient for the term corresponding to this interaction effect should be positive in sign.[6] When C represents the criterion and A and B represent the two predictors, the equations fitted in the two steps are:

$$C = b_0 + b_1(A \cdot B) + e \quad (1)$$

$$C = b_0 + b_1(A \cdot B) + b_2(A) + b_3(B) + e \quad (2)$$

In these equations, the four regression coefficient values are the constant or intercept term (b_0), the interaction effect (b_1), and the effects of Predictors A (b_2) and B (b_3); A·B is the interaction predictor variable, which is formed by multiplying values of A and B. The usual procedure for testing an interaction effect is to enter the variable representing it into the regression analysis *after* estimating main effects by entering the interaction's individual component variables as predictors (Cohen & Cohen, 1983). However, to determine if the data of the balanced identity design can be fit entirely by the interaction term this typical order must be reversed.[7] Good fit for the interaction-only model of Equation 1 will appear as the absence of a statistically significant increment in *R* on Step 2.

Stated more completely, Prediction 1 leads to four expectations for results of the 2-step hierarchical analysis: (a) The *R* in Step 1 should account for substantial variance in the criterion and should estimate a numerically positive value for b_1, (b) the estimate of b_1 should also be positive in Step 2,[8] (c) the increment in *R* on Step 2 should not be statistically significant, and (d) neither b_2 nor b_3 should differ significantly from zero in Step 2. The last two predictions require a scaling assumption – that numeric zero values for variables A and B indicate zero strength of the associations that they measure. Failure of this scaling assumption could produce a significant increment of the multiple *R* in Step 2, along with significant deviations of b_2 and/or b_3 from zero (Aiken & West, 1991, Appendix A). Therefore, when Predictions (a) and (b) are confirmed, minor failures of (c) and (d) could be due to inadequacy of the scaling assumption rather than being due to invalidity of Prediction 1.

---

[6]To explain: Figure 6 shows that the criterion measure is expected to be especially low when one predictor is high and the other is low (a situation that makes the AB product negative). Similarly, Figure 6 shows that the criterion is expected to be especially high either when both predictors are low or when both are high — both of these are situations that make the AB product positive. With the value of the criterion therefore expected to be especially low when the AB product is negative, and especially high when the AB product is positive, the overall expectation is a positive relation between the AB product and the criterion.

[7]We are grateful to an anonymous reviewer for suggesting this analysis strategy.

[8]As implied by Figure 6, when there is no variation on one predictor, the regression analysis can degenerate to a linear relationship between the other two variables, with no interaction effect. Such an extreme circumstance should be rare. Nevertheless, this reasoning indicates that when the range of at least one variable in the balanced identity design is restricted, the interaction term in Step 2 may be only weakly positive, and therefore not necessarily statistically significant.

Zero-order correlations. Zero-order correlations are unadjusted product-moment correlations between two variables. These are distinct from the partial correlations that are produced in a multiple regression analysis. As explained a few paragraphs previously, the expected zero-order correlation between any two of the three variables in a balanced identity design depends on the distribution of the third variable. If Figure 6's Predictor B is well distributed across its full range, the zero-order correlation between A and B should be a mixture of the three slopes shown in Figure 6, with no directional expectation — it should not differ significantly from zero. The situation is different when Predictor B has a *polarized* distribution, meaning that scores are noticeably displaced to one or the other side of zero. For example, consider a balanced identity test of the analysis shown in Figure 5. In this design, the three measures of the balanced identity design are self-female association (group identity), self-positive association (self-esteem), and female-positive association (group attitude). In a study that obtains these three measures for women subjects, the subjects should have scores polarized toward high values on the group identity measure (that is, women subjects should associate self much more with female than with male). If this self-female association measure is in the role of Figure 6's Predictor B, the obtained data for the other two variables should fall near Figure 6's positive regression slope that is labeled 'High Predictor B'. That is, the zero-order correlation between the measures of self-positive and female-positive associations should be numerically positive. If the sample were instead polarized toward the low end of Predictor B (e.g., a sample of men), the correlation of Predictor A with Criterion should be negative, corresponding to the slope labeled 'Low Predictor B'.

To summarize and generalize: When any variable in the balanced identity design is polarized toward its high end, the zero-order correlation between the other two variables should be positive; when any of the variables is polarized toward its low end, the zero-order correlation between the other two variables should be negative; and if a variable in the balanced identity design is not polarized, correlations between the other two variables should not differ from zero.

Illustrative Experiment:

Balanced Identity Investigation of Gender Attitude

The first experiment that used a balanced identity design was one that was designed to test Prediction 1 in the domain of gender identity and gender attitude, as illustrated in Figure 5.[9] This test required (a) measuring three types of associations: self-gender (gender identity), self-valence (self-esteem), and gender-valence (gender attitude), and then (b) testing the regression relationships specified by Prediction 1. For the IAT, it was necessary to represent the gender, self, and valence concepts of Figure 1 in the form of contrasts of complementary categories. Thus, *self* was represented by the contrast of self vs. other; *valence* by the contrast of pleasant vs. unpleasant, and *gender* by the contrast of male vs. female.

[9]This experiment was reported by Farnham and Greenwald (1999). A detailed description can be found in Farnham (1999, Experiment 1).

Subjects

The participants were 67 undergraduate women at University of Washington who volunteered in exchange for a small amount of extra credit in their introductory psychology courses. Data for two subjects were excluded for not following instructions. An additional 8 subjects had incomplete questionnaire data, leaving N = 57 for analyses of explicit measures.

Procedure

Subjects participated individually. The procedure consisted of administering both explicit (self-report, paper-pencil) and implicit (IAT, computer-administered) measures of the three sets of associations that constituted the balanced identity design. When this experiment was conducted it was suspected that completing the IAT measures was more likely to influence responses to the self-report measures than vice versa. (However, there are not yet any data that establish systematic effects of order of administering IAT and self-report measures.) Consequently, the self-report measures of association strengths were administered first.

Explicit Measures

Self-esteem. Explicit self-esteem (association of self with positive valence) was measured with three procedures: (a) a thermometer measure, (b) a Likert measure, and (c) a standard self-esteem inventory. For the thermometer measure, subjects rated both 'yourself' and 'other people' by placing a horizontal mark through a vertical thermometer scale that was anchored '0 – Cold or unfavorable' at the bottom, '50 – Neutral' in the middle, and '99 – Warm or favorable' at the top. The measure was constructed as a difference score, subtracting the score for 'other people' from that for 'yourself'. For the Likert measure, subjects rated each of 6 pleasant-meaning and 6 unpleasant-meaning words on 7-point scales that were anchored at their ends by 'not at all characteristic of you' (1) and 'extremely characteristic of you' (7). (The 12 items are listed below in describing the IAT measures.) The measure was constructed by subtracting the average score for the 6 unpleasant items from that for the 6 pleasant items. The standard inventory measure was the 10-item Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965). The balanced identity statistical analyses required each of these measures to be scored on a scale that had a rational zero point. The thermometer and Likert measures had rational zero points due to their construction as difference scores. The RSES asks subjects to respond to positive and negative self-descriptive statements on a 4-point agreement scale, and could therefore be given a rational zero point by letting zero correspond to the midpoint (between the 2nd and 3rd points) of the agreement scale. An overall explicit self-esteem measure was obtained for each subject by first dividing each of the three self-valence measures by its standard deviation and then averaging the three values. (This procedure preserved the desired location of the zero point.)

Gender identity. Explicit gender identity (self-gender association) was measured in Likert format. Subjects rated each of 6 male and 6 female nouns (listed below in describing IAT measures) on a 7-point scale anchored by 'not at all characteristic of you' (1) and 'extremely characteristic of you' (7). This measure was scored by subtracting the average score for the 6 male items from that for the 6 female items – as a consequence high scores represented stronger association of self with female.

Gender attitude. The explicit gender attitude variable combined two measures, one in thermometer format and one in Likert format. The thermometer measure was like that for self-esteem – the two concepts that were rated on the warmth-of-feeling scale were 'females' and 'males'. The difference score for this measure used the rating for 'males' subtracted from that for 'females'. The Likert measure of gender attitude was also parallel to the one for self-esteem, being constructed from the subject's ratings of 6 pleasant and 6 pleasant items twice each, once on a scale ranging from 'not at all characteristic of males' (1) to 'extremely characteristic of males' (7) and once on a similar scale referring to 'females'. Attitude scales for both concepts were obtained by subtracting the average score for the 6 unpleasant items from that for the 6 pleasant items. The gender attitude difference score was then computed as the attitude score for males subtracted from that for females. A combined gender attitude measure was constructed for each subject by first dividing each of the two gender-valence association measures by its standard deviation, then averaging the resulting two values. Higher scores represented stronger association of female with positive valence.

IAT Measures

Subjects completed three computer-administered IAT procedures that yielded measures of implicit self-esteem, gender identity, and gender attitude. These IAT measures were as parallel as possible to the Likert explicit measures of each of these three constructs, achieved by using the same stimulus items for both types of measure. The order of the three IAT measures was counterbalanced across subjects. After an initial analysis revealed that order of administration of the IAT measures had no systematic effect, order was dropped as a predictor in subsequent analyses.

Each of the three IAT measures used the five-step schema shown in Figure 3. Each step involved a block of 20 trials that was treated as practice. Steps 3 and 5, the two tasks that provided data for the IAT measure, each had an additional 40-trial block of data collection trials. Mean performance measures for the two combined tasks were computed using procedures described by Greenwald et al. (1998). These included (a) dropping the first two trials of the 40-trial blocks because of their typically lengthened latencies, (b) analyzing latencies for all trials, including those on which errors were made, (c) recoding latencies below 300 ms to 300 ms and those above 3000 ms to 3000 ms, and (d) log-transforming the resulting data before computing average performances. As pointed out by Greenwald et al. (1998), these procedures eliminate some statistical noise, but do not produce results that are substantively different from those obtained with several reasonable alternative data-management procedures. IAT scores were computed as the difference in means between the two combined tasks, always scored in the same direction as previously described for the Likert measures.

Idiographic self and other items. Prior to any of the IATs, each subject was asked to generate an item representing *self* in response to each of seven cues: first name, middle name, last name, home city, home state, home country, and race/ethnicity.[10] Subjects entered each item by typing it in a dialog box on the computer display screen. After generating the seven items, subjects selected an additional item in each of the same seven categories to represent *other*. For each of the 'other' selections, subjects were provided with a wide range of choices for each probe and were asked to select, for each, one of these that was neither associated with themselves nor was specially liked or

[10]This differs from the pronoun, or *generic*, representation of self and other shown in Figure 3. Greenwald and Farnham (2000) used both of these formats on the same subjects, finding that the resulting measures correlated well with each other and had similar correlations with third variables..

disliked. Lastly, subjects were given the opportunity to drop one or two items from each of the 7-item self and other categories if, in retrospect, the generated or selected items did not well represent those categories.

Self-esteem. In addition to items for the self-other contrast, the self-esteem IAT required items for a pleasant-unpleasant (valence) contrast. For the valence contrast, pleasant was represented by 6 words (joy, warmth, gold, happy, smile, pleasure) and unpleasant by 6 words (gloom, agony, pain, stink, filth, death). (These were the same12 words used in the Likert measures of self-esteem and gender attitude.) The self-esteem IAT score was the difference score computed by subtracting mean performance in the block for which the task was to classify self-or-pleasant vs. other-or-unpleasant from that in the block for which the task was to classify other-or-pleasant vs. self-or-unpleasant. High scores therefore represented association of self with positive (more than negative) valence.

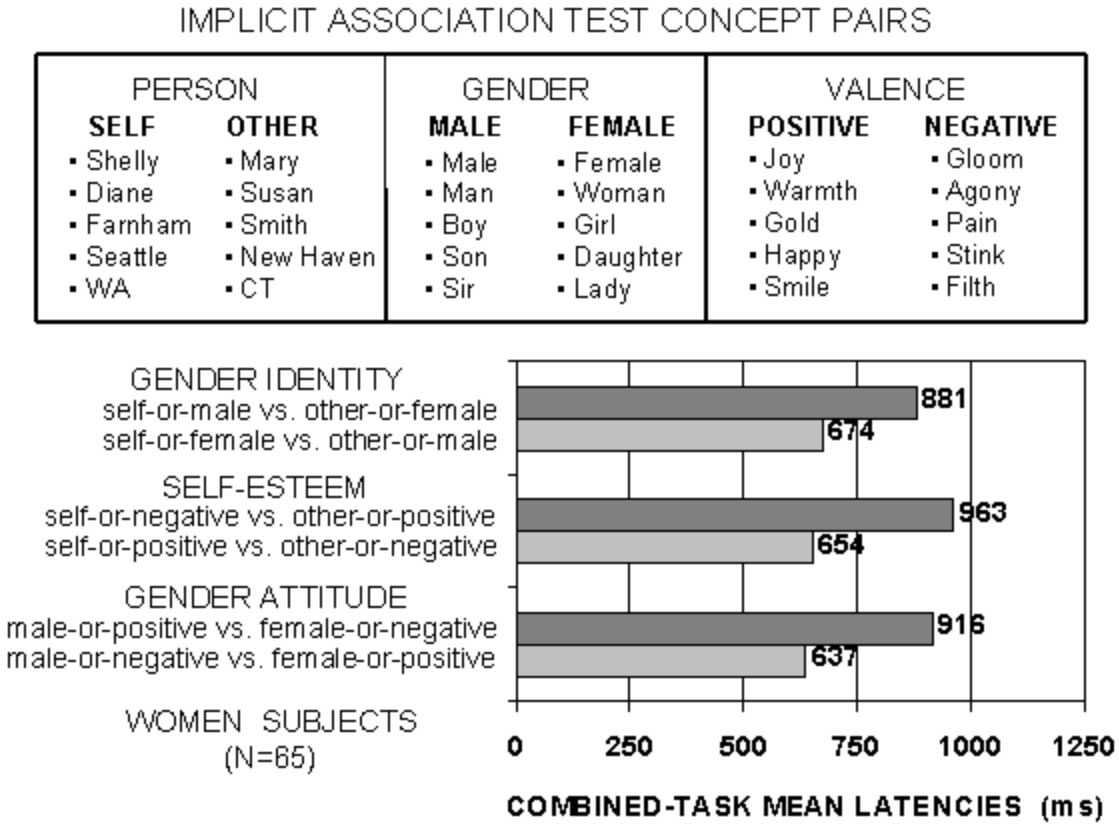

Figure 7. **IAT stimuli and measure from gender attitude experiment**. Items shown in the upper part of the figure represent most of those used in the three IAT measures. The items shown for *self* and *other* are ones that might have been selected had one of this article's authors (SDF) been a subject. The data graph presents mean latencies for the two combined tasks included in each of the three IAT measures. The mean IAT effect for each measure is the upper-bar-minus-lower-bar difference between the means shown for each combined task. (Data from Farnham & Greenwald, 1999)

Gender identity. The gender identity IAT used the same self-other contrast that was used for the self-esteem IAT. For the gender (male-female) contrast, male was represented by 6 words (man, boy, son, sir, guy, male) and female by 6 parallel words (woman, girl, daughter, madam, lady, female). (These were the same 12 words used in the Likert measures of gender identity and gender attitude.)

The difference score for the gender identity IAT measure was computed such that higher scores represented greater association of self with female than with male.

Gender attitude. The gender attitude IAT used the same male-female contrast as the gender identity IAT and the same pleasant-unpleasant contrast as the self-esteem IAT. Its difference score was computed such that higher scores represented greater association of female than male with positive valence.

Results and Discussion

Figure 7 summarizes the three IAT measures and presents mean latencies for the two combined tasks of each IAT measure (those corresponding to Steps 3 and 5 in Figure 3). Figure 8 presents data distributions for all of the IAT and self-report measures.

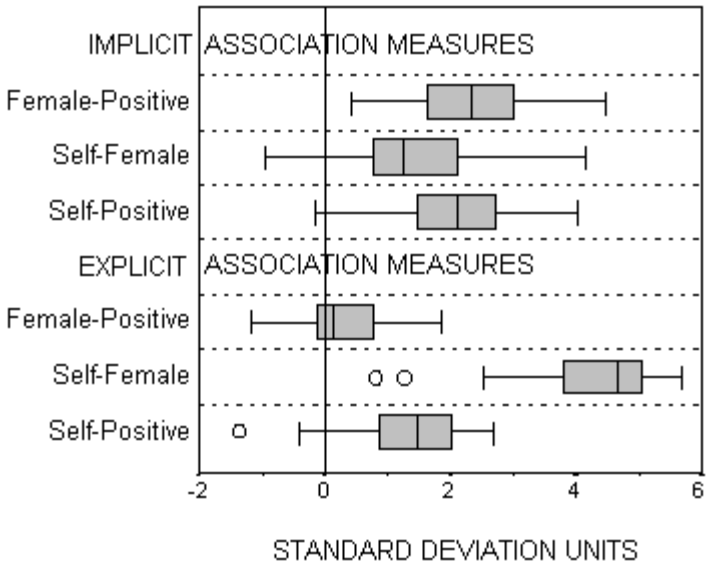

Figure 8. **Distributions of implicit and explicit measures for illustrative balanced identity design.** These boxplots show ranges, medians, quartile boundaries, and the few outlier cases (circles) for the implicit (IAT) and explicit (self-report) measures of associations in the illustrative experiment. (Women subjects: N = 65 for implicit measures; N = 57 for explicit measures; data from Farnham & Greenwald, 1999.)

Implicit measures – descriptive results. The upper half of Figure 8 shows that all three of the implicit measures were polarized toward high values. This was not surprising. The sample of women college students was expected, on average, to display ingroup identity as female (i.e., polarized high scores for self-female association) as well as positive scores on self-esteem that are typical for student samples (polarized high scores for self-positive association). The polarized high scores on the third measure (female-positive association) were then expected from Prediction 1. That is, as can be seen in Figure 6, when two predictor variables both have high values, the third (criterion) variable is also expected to have high values. Because all three measures were polarized toward high values, all three zero-order correlations were expected to be positive. In Figure 9, these zero-order correlations are presented on the sides of the inner triangle in the left panel. Consistent with expectation, all three of these correlations were positive ($ps \leq .01$).

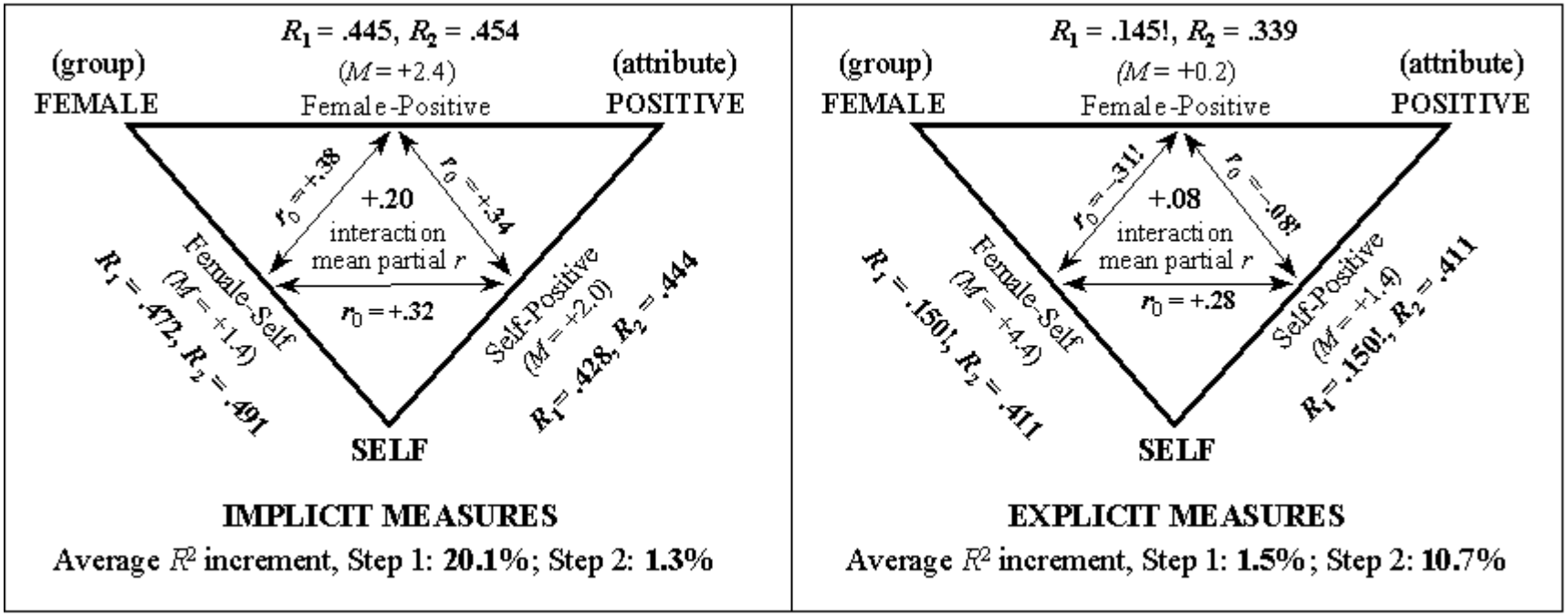

Figure 9. **Summary of statistical tests for implicit and explicit measures**. The illustrative experiment used a balanced identity design with implicit (left panel) and explicit (right panel) measures of associations among the self, group, and attribute concepts that are named at the corners of each outer triangle. All mean values (*M*) are reported in SD units, with location of zero untransformed. Each zero-order correlation (r_0) on an edge of an inner triangle relates the two association measures pointed to by the contiguous edge's two arrows. These r_0s are consistent with Prediction 1 of the present theory if they have the same sign as the mean value of the measure of the remaining association, which is shown on the parallel (i.e., opposite) edge of the outer triangle (see text discussion of Figure 6). For example, the r_0 between female-positive and self-positive association measures was expected to be positive in both panels because the mean value for the remaining association, female-self, was positively polarized in both. The correlation of +.34 in the left panel is therefore theory consistent, but the corresponding correlation of –.10 in the right panel does not have the predicted positive sign and is therefore inconsistent with theory. The mean partial *r* in each inner triangle is the average of the interaction effect partial *r*s obtained in the second steps of the three hierarchical multiple regression analyses that are summarized in each panel. R_1 and R_2 identify the multiple regression coefficients produced, respectively, in the first and second steps of these regression analyses. Exclamation marks (!) follow r_0 values that are opposite in sign from prediction and R_1 values that are associated with opposite-from-predicted (i.e., negative) interaction effect coefficients. For either r_0 or R_1 in the left panel (N=65), values of .244, .317, and .344 are associated, respectively, with p = .05, .01, and .005. For either r_0 or R_1 in the right panel (N=57), values of .261, .338, and .367 are associated with p = .05, .01, and .005. (Data from Farnham & Greenwald, 1999.)

Implicit measures – hierarchical regressions. The three regression tests of Prediction 1 – one using each of the three IAT measures as criterion – are summarized in the left panel of Figure 9. In Figure 9, R_1 is the regression coefficient from the first regression step, in which only the interaction term was entered as a predictor. R_2 is the coefficient from the second step, which added the interaction's two component variables as predictors. These two regression steps correspond to previously presented Equations 1 and 2.

As previously described, each of the three regression analyses provided four indicators of fit with Prediction 1. To review, the four indicators are: (a) in Step 1, a substantial R_1 associated with a positive value of b_1, (b) a positive value of b_1 also in Step 2, (c) a nonsignificant increase from Step 1 to Step 2 in criterion variance explained, and (d) neither b_2 nor b_3 significantly different from zero in Step 2. All four of these indicators appeared as predicted in each of the three regression analyses summarized in the left panel of Figure 9. In particular: (a) The three standardized values for b_1 in Step 1 averaged +.449 (ps $\leq$ .0004); (b) the b_1 coefficients in Step 2 were all positive in sign, with partial rs averaging +.20; (c) the increments in R^2 from Step 1 to Step 2 were all nonsignificant (ps $\geq$ .48), and (d) all b_2 and b_3 values in Step 2 were nonsignificant (ps $\geq$ .25). In sum, these results were unequivocally consistent with the present theory's Prediction 1.

Explicit measure results. The inner triangle of the right panel of Figure 9 indicates that all three zero-order correlations among the explicit measures differed from Prediction 1's expectations based on distributions for the three measures (see lower half of Figure 8). Two zero-order correlations that were expected to be positive in sign were negative, and the third, which was expected to be near zero, had a statistically significant positive value. For the multiple regression results (also summarized in the right panel of Figure 9) the exclamation marks after each of the three R_1 values indicate that the b_1 coefficients associated with these R_1s were all opposite from prediction in sign (i.e., negative), clearly not conforming to Prediction 1. The failure of Prediction 1 at Step 1 of the regression analysis made the Step 2 results irrelevant to evaluating Prediction 1.

In summary, results of the illustrative experiment's data conformed well to the present theory's Prediction 1 for implicit measures, but not for explicit measures. For implicit measures, Equation 1 explained a substantial average of 20.1% of criterion variance, with very little additional variance (average of 1.3%) explained by Equation 2. For the explicit measures, what little variance was explained by Equation 1 (average of only 1.5%) was directionally inconsistent with Prediction 1.

Additional Tests of Prediction 1

Banaji, Nosek, Greenwald, and Rosier (1999). Banaji et al. (1999) used a balanced identity design in a study of racial identity, racial attitudes, and self-esteem. This design replaced the gender (male-female) contrast of Figure 7's illustrative experiment with a race (Black-White) contrast. The subjects were undergraduate male and female students at Yale University – 30 African American (Black) and 31 European American (White). Because Banaji et al.'s balanced identity design was complete only for implicit measures, explicit-measure results are not described here. The data from regression analyses of their implicit measures conformed well to the first two of Prediction 1's four expectations: (a) the three R_1 coefficients were all positive, accounting for a very substantial average of 29.9% of criterion variance; and (b) the b_1 coefficients in Step 2 were all positive in sign, with partial rs averaging +.35. The remaining two tests revealed mild deviations from Prediction 1: (c) although the increments in R^2 from Step 1 to Step 2 were, as expected, considerably smaller than for Step 1– averaging 7.2% of additional explained variance – two of the three were statistically

significant, and (d) one of the total of six b_2 and b_3 values at Step 2 differed significantly from zero. (Additional details of method and data can be found in the Appendix.)

Mellott and Greenwald (2000). Mellott and Greenwald (2000) used a balanced identity design to investigate relationships among age identity, ageist attitudes, and self-esteem. Their subjects were 52 college students (mean age = 19.7, SD = 1.6) and 46 older subjects (mean age = 74.7, SD = 6.6). Descriptively, there was a surprise in finding that the implicit age identity measure (self-old association) was polarized toward its low (self-young) end. This was surprising because inclusion of both young and older subjects was expected to produce a distribution centered near zero for that measure. However, the older subjects showed approximately the same implicit identification with young and the same implicit negativity toward old age that younger subjects did. For the implicit measures, Step 1 regression results were as expected, with numerically positive and statistically significant b_1 values accounting for an average of 13.1% of criterion variance. For explicit measures, by contrast, two of Step 1's b_1 values were numerically negative and none was statistically significant, account for an average of only 1.0% of criterion variance. For Step 2 of the implicit measure regression analyses, results did not fit very well with Prediction 1. The average of the three b_1 values in Step 2 was very near zero and two of the three regressions revealed both statistically significant increments in criterion variance explained at Step 2 (average of 4.2%) and statistically significant deviations from zero values for coefficients of individual variable predictors (i.e., b_2 and/or b_3). In summary, Mellott and Greenwald's implicit measure data showed partial support for Prediction 1, whereas their explicit measure data showed no support. (Additional details of method and data are in the Appendix.)[11]

Balanced Identity and Stereotypes

Prediction 1 applied the balance-congruity principle to the triad of ingroup, self, and positive (valence). To the extent that both self-positive and self-group associations exist (a configuration with a shared first-order link), a group-positive association was expected to develop. The next step in exploring and testing the present theory was to apply the balance-congruity principle to configurations in which the positive valence attribute of Prediction 1 is replaced with a trait attribute. This substitution creates a trio of associations consisting of an identity, a stereotype, and a self-concept (see Figure 4). When self is associated with a group that is stereotypically associated with a trait, there is a shared first-order link that connects both self and trait to the (in)group. The balance-congruity principle then predicts strengthening of the link between self and trait. For example, consider the stereotypic association of *female* with the trait of *warmth* (nurturance). If a woman associates self with female and female with warmth, then she should also associate self with warmth.

Prediction 2: **Balanced identity and stereotype**. *Strength of an association between self and a trait attribute should be a multiplicative function of the strengths of associations of self to group (ingroup identity) and of group to attribute (group stereotype).*

[11] A replication of Mellott and Greenwald's (2000) study was recently reported by Hummert, Garstka, O'Brien, Savundranayagam, & Zhang (2000). Examination of their findings using the hierarchical regression method matched the result of Mellott and Greenwald (2000), being consistent with Prediction 1 at Step 1 for implicit measures, but not for explicit measures, and not fitting very well with prediction at Step 2.

Test of Prediction 2 with Gender Stereotypes of Warmth and Potency

Prediction 2 was testable using data that had been obtained in an investigation of implicit gender stereotypes concerning potency and warmth (Rudman et al., in press). The gender-trait design of Rudman et al.'s Experiment 4 differed from the gender-attitude design of Figure 7 by using the trait attribute contrast of potency versus warmth in place of the positive versus negative valence contrast. In this design, Prediction 2 translates to the expectation (worded from the perspective women) that strength of the association of self with warmth should be a joint function of the strength of gender identity as female and strength of the gender-stereotypic association of female with warmth.

In Rudman et al.'s (in press) Experiment 4, the stereotypically male-associated attribute of potency was represented in IATs by the words, *power*, *strong*, *confident*, *dominant*, *potent*, *command*, and *assert*. The stereotypically female-associated attribute of warmth was represented by *warm*, *nurture*, *nice*, *love*, *caring*, *gentle*, and *kind*. Subjects were 43 undergraduate men and 52 undergraduate women at University of Washington. The explicit measure of stereotype was obtained by having subjects rate each of the IAT's words representing potency and warmth on two separate 7-point scales, one each assessing the word's accuracy as a description of *men* and *women*. Similarly, the self-concept measure involved rating all of these same attribute words on two 7-point scales, one each assessing the word's accuracy as a description of *self* and *others*. Each measure was computed as a difference score such that zero values indicated that potency and warmth were equally applicable to the two contrasted concepts (men and women for the stereotype measure; self and others for the self-concept measure). The explicit gender identity measure was obtained from a 4-item Likert-format scale, scored so that high scores indicated self-identification as female.

Just as for Prediction 1, it was possible for any of the three associations (female-self, female-warm, self-warm) to be conceived as dependent on values of the other two. Therefore, the statistical strategy was to use the same multiple regression format as for Prediction 1, conducting a 2-step hierarchical analysis with each of the three associations, in turn, in the role of criterion variable. These tests of Prediction 2 are summarized in Figure 10.

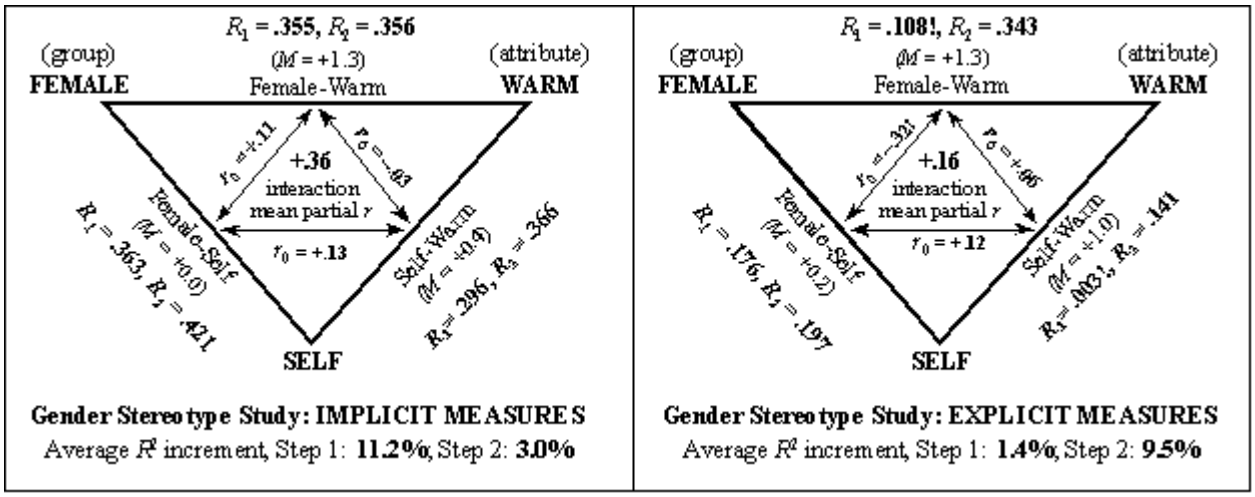

Figure 10. **Summary of balanced identity multiple regression analyses**. This summary uses the format of Figure 9. The implicit measure data were fully consistent with the present theory's Prediction 1 in both Steps 1 and 2 of all three multiple regression analyses. By contrast, the explicit measure data were markedly inconsistent with Prediction 1 at Step 1, making the Step 2 results irrelevant. For either r_0 or R_1 in both panels (N=95), values of .202, .263, and .286 are associated, respectively, with p = .05, .01, and .005. (Data from Rudman et al., in press, Experiment 4.)

Test of Prediction 2 with implicit measures. As expected for a sample that included both women and men, the implicit gender identity measure (self-female association) had a nonpolarized distribution. The self-concept measure was slightly polarized toward high values, indicating an average tendency to associate self more with warmth than with potency. The only clearly polarized implicit measure was the gender stereotype measure (female-warmth association), which was polarized toward high values, consistent with the expected stereotypical association of female more with warmth and of male more with potency. With polarization occurring only on the gender stereotype measure, the only zero-order correlation that was expected to differ from zero was that between self-concept and gender identity. Although that correlation was positive as expected (r = +.13, p = .21) it, along with the other two zero-order correlations, did not differ significantly from zero.

Effects expected from Prediction 1 were very clearly apparent in most of the results of the three 2-step multiple regression analyses. The three standardized values for b_1 in Step 1 were all positive, averaging +.338 (ps $\leq$ .004). The b_1 coefficients in Step 2 were all positive in sign, and all three Step-2 interaction partial rs were statistically significant (ps $\leq$ .001), averaging +.36. Lastly, the increments in R^2 from Step 1 to Step 2 were all nonsignificant (ps $\geq$ .09). The only weaknesses in support for Prediction 1 were that the percentage of criterion variance explained by the Step 1 model was only moderate (average of 11.2%) and two of the three regressions each had one statistically significant individual-variable predictor in Step 2. As explained previously, these Step-2 deviations from prediction could be due to a failure of the assumption that numeric zero on the implicit measures corresponded to absence of association.

Test of Prediction 2 with explicit measures. Two of the explicit measures were polarized – gender stereotype (associating female more with warmth, male more with potency) and trait self-concept (associating self more with warmth than potency). This led to the expectation of two positive zero-order correlations, only one of which was positive (gender identity with trait self-concept; r = +.12, p = .25, see Figure 10). The other, contrary to prediction, was significantly negative (gender identity with gender stereotype; r = –.32, p = .002). The most critical test of Prediction 1 is from Step 1 of the three hierarchical regression analyses. As can be seen in Figure 10, two of the three regressions lacked the predicted positive coefficient for the b_1 coefficient in Step 1 and the one positive b_1 coefficient in Step 1 was weak (R_1 = .176, p = .09). With this lack of support for Prediction 1 in Step 1, the results for Step 2 were irrelevant to Prediction 1.

Additional Test of Prediction 2 with Math-Gender Stereotype

Nosek, Banaji, and Greenwald (2000) investigated self-concepts, gender stereotypes, and attitudes toward academic subject areas in a sample of 46 male and 45 female Yale University undergraduate students. The two academic domains that were contrasted in their implicit and explicit measures were *mathematics* and *arts*. (Additional details are provided in the Appendix.) The balanced identity design was expected to reveal that association of self with mathematics would be consistent with the combination of one's gender identity as male or female and one's possession of the gender stereotype that associates mathematics with male.

This experiment had a balanced identity design only for implicit measures. The multiple regression results of these implicit-measure data were fully consistent with expectations based on Prediction 1. The three standardized values for b_1 in Step 1 were all positive, averaging +.359 (ps $\leq$ .03). The b_1 coefficients in Step 2 were all positive in sign, averaging +.16 (one was statistically significant). Additionally, the increments in R^2 from Step 1 to Step 2 were all nonsignificant (ps $\geq$

.30) and there were no significant deviations from zero values for the interaction-component predictors in Step 2. The interaction-effect-only analysis of Step 1 explained an average of 13.5% of criterion variance, with Step 2 adding an average of only 1.5%. A description of these implicit-measure findings from the perspective of women is that strongly gender-identified women who have the stereotypic male-math association are unlikely to associate self with math. The stereotype therefore becomes an obstacle to women's career aspirations in math. From the male perspective, the combination of gender identity and stereotype become factors that support an association of self with math.

Corollary of Prediction 2: Positive Valence of Ingroup-Stereotypical Traits

Prediction 2 relates the strength of a component of self-concept (a self-trait association) jointly to the strengths of an ingroup identity (self-group association) and a relevant stereotype (group-trait association). The possible participation of this self-trait link in further balance-congruity effects yields a corollary of Prediction 2. The self-trait link, together with a self-positive link (self-esteem), produces a configuration with a shared first-order link in which both the trait and positive valence are associated with self. Consider, as an example, women who (a) associate self with warmth (which is expected from Prediction 2, due to associating self with female and female with warmth), and (b) also associate self with positive (i.e., have high self-esteem). This combination creates a shared first-order link of both *warmth* and *positive* to *self*. The balance-congruity principle, applied to this configuration, predicts strengthening of the association of *warmth* with *positive*. The corollary of Prediction 2 describes this second-order operation of the balance-congruity principle.

> **Corollary of Prediction 2**: *An attribute that is stereotypically associated with one's ingroup should acquire positive valence.*

According to the corollary, the attitude (valence associated with the ingroup-stereotypic trait) should be a joint (multiplicative) function of strengths of the self-group (ingroup identity) and group-trait (stereotype) associations. This is expected because strength of the self-trait link is expected to be described by that multiplicative function.[12] The corollary can therefore be tested with the same 2-step multiple regression format used for Predictions 1 and 2. Further, because the corollary treats attitude as a consequence of the prior existence of ingroup identity and group stereotype, it is reasonable for this test to use just the single hierarchical regression in which the attitude measure (trait-valence association) is criterion.

Test with math-gender attitude. The Nosek et al. (2000) experiment included an implicit measure of math-valence association, which made it possible to test Prediction 2's corollary. Worded from the perspective of men, the corollary predicts that positive attitude toward math should be a joint function of the strengths of male identity (self-group association) and the gender-stereotypic association of mathematics with male. Results of the 2-step regression agreed well with Prediction 2's corollary. The standardized value of b_1 in Step 1 was positive, +.359 ($p = 10^{-5}$). The b_1 coefficient in Step 2 was also positive with a partial r of +.30 ($p = .004$). Also as expected, the increment in R^2 from Step 1 to Step 2 was nonsignificant ($p = .50$) and neither of the individual-variable predictors' coefficients differed from zero when entered in Step 2 (ps $\geq .30$). Equation 1 explained 20.2% of variance in the implicit math attitude measure, and Equation 2 added only an additional 1.3%. These

[12] An additional assumption on which the prediction depends is that self-esteem is both positive and not greatly variable across subjects in the research sample. I.e., the test assumes that individual differences in the self-trait link are much more substantial than individual differences in self-esteem.

results unequivocally supported Prediction 2's corollary.[13] A description of the result is that both men's liking and women's disliking for math are magnified by the strengths of their implicit gender identities and their implicit gender stereotypes.

Discussion of Empirical Findings

Empirical Summary

All predictions from the present theory have been stated as a dependence of the strength of one (criterion) measure of association on the multiplicative product of the strengths of two other (predictor) measures of association among the three conceptual elements of the balance triad. These theory-based predictions were tested using 2-step hierarchical regression analyses in which only the multiplicative product of the two predictor measures was entered on the first step, and the individual predictor measures were added on the second step. This article includes sixteen such analyses for implicit measures and nine for explicit measures. The most important expectation for these analyses was that (a) the multiplicative product term would have a statistically significant positive regression coefficient in Step 1. Strikingly, this expectation was confirmed for all sixteen of the implicit measure analyses, while being confirmed for none of the explicit measure analyses.

After finding of the significant positive coefficient for the multiplicative product term in Step 1, it was appropriate (i.e., for the implicit measures) to examine three additional expectations concerning results from the second step of the regression analysis: (b) that the coefficient of the multiplicative product term would remain positive in Step 2 – this was confirmed for 15 of the 16 implicit measure tests, (c) that Step 2 would not add significantly to the variance accounted for by Step 1 – confirmed for 12 of 16 tests, and (d) that coefficients for the individual predictor variables would not differ significantly from zero in Step 2 – confirmed for 11 of 16 tests. As explained previously, it is not presently possible to determine whether the nonconfirmations for the last two expectations should be interpreted as damaging to theory or, alternately, attributed to failure of the assumption that values of zero on the implicit measures could be interpreted as absence of association.

Why Did Consistency Appear Only on Implicit Measures?

The present theory's triadic consistency predictions were largely confirmed in the data for implicit measures. By contrast, there was no evidence of such consistency in the data for explicit measures. This combination of observations amounts to an empirical dissociation between the two types of measures, and it poses a straightforward question: Why? What difference between implicit and explicit measures can explain the occurrence of predicted findings only for the implicit measures?

The original rationale for using implicit measures in the present research was twofold: (a) *introspective limits* – implicit measures might be able to measure associations for which the respondent lacks awareness, and (b) *response factors* – self-report of associations of which the respondent is aware might be masked by factors such as demand characteristics (Orne, 1962), evaluation apprehension (Rosenberg, 1969), and subject role playing (Weber & Cook, 1972). These two interpretations are complementary in the sense that one of them assumes a set of influences (unconscious knowledge) that affect implicit, more than explicit, measures, whereas the other assumes influences (response factors) that primarily affect explicit measures. The introspective limits

[13] Tests with the other two measures as criterion also showed results that were fully consistent with Prediction 1.

and response factors interpretations are not mutually contradictory – the implicit-explicit dissociation might be explained as well by assuming their joint operation as by assuming their individual operation.

Common to both the introspective limits and response factors interpretations is an assumption that the IAT provides better access to associative knowledge than does self-report. This may appear surprising to those who interpret cognitive consistency as a rational process that, by virtue of operating in a conscious arena, should be fully accessible to self-report. However, there exists no body of research to support the assumption that cognitive consistency is a consciously imposed attribute of knowledge structures. The question of conscious versus unconscious operation of consistency processes was simply not an issue for cognitive consistency theories of the 1950s. Their neglect of what appears to be an important theoretical question can be understood by appreciating that, until quite recently, the vast majority of social (and other) psychologists tacitly assumed that cognitive processes were generally conscious and therefore available to self-report. It was only after Nisbett and Wilson's (1977) attack on this tacit introspectionism that social psychologists began to take seriously the possibility that significant portions of social knowledge might be inaccessible to awareness.

What Does the IAT Measure?

Related to the two accounts of implicit-explicit dissociation are two interpretations of how the IAT differs from self-report in providing a measure of association strength. If one assumes that association strengths measured by the IAT are consciously accessible, then the difference between self-report and IAT measures can be understood as the difference between direct and indirect measures of association strength. In this view, the IAT works well as an indirect measure because of its presumed lack of susceptibility to response factors that affect direct measures. However, if one assumes that association strengths measured by the IAT are often not accessible to conscious inspection, then the difference between self-report and IAT can additionally be understood as a difference in their access to unconscious knowledge. Again, there is no need to assume that only one of the two explanations is correct.

A possible means of appraising the relative roles of introspective limits and response factors is to identify conditions under which dissociations between implicit and explicit measures do and do not occur. Some progress in that direction has been made by identifying domains in which dissociations between IAT and self-report measures occur, and others in which agreement between the two types of measures is found. Dissociations have been identified in studies of attitudes toward gender (Greenwald & Farnham, 2000), race (Banaji et al., 1999; Greenwald et al. 1998), ethnicity (Greenwald et al., 1998), and age (Mellott & Greenwald, 2000), and in the domain of gender stereotypes (Nosek et al., 2000; Rudman, Greenwald, & McGhee, in press). On the other hand, agreement between implicit and explicit measures has been observed in studies of IAT-measured and self-report-measured attitudes toward political candidates (Nosek, Banaji, & Greenwald, in press) and toward some consumer products (Brunel, Collins, Greenwald, & Tietje, 1999; Maison, Greenwald, & Bruin, in press). The domains characterized by high correlations are ones for which it is plausible that subjects have little motivation to disguise their attitudes on explicit measures. Accordingly, the non-dissociation results fit with a response-factors interpretation of the implicit-explicit difference. Unfortunately, it seems almost equally plausible that the domains in which non-dissociation results occur are ones that afford superior introspective access to associations. For the present, the distinction between the introspective-limits and response-factors interpretations of the difference in patterns of results for implicit and explicit measures is difficult to resolve.

Neglected Causation

The present theory's Principle 1 (balance-congruity) describes a causal effect of the shared-first-order-link configuration on the strength of association between the two nodes that are associated to the same third node. By contrast, Predictions 1 and 2 – both of which were based on Principle 1 – were stated in terms of correlations among contemporaneous strengths of associative links among three concepts, ignoring causation. Likewise, the empirical tests provided by the balanced identity design are silent on causation. This neglect of causation was unavoidable in testing Principle 1 in natural cognitive structures. For these structures, one might guess at the order in which the various associations were formed, but the data of the balanced identity design cannot evaluate those guesses. Efforts to evaluate the causal content of the balance-congruity principle will likely require studies of experimentally created concepts and associations among them.

Additional (Untested) Predictions

Resisting Association to Both of Two Bipolar-Opposed Nodes

Because members of demographically diverse groups within a society often have broadly shared cultural experience, the associative base of social knowledge for members of distinct groups may be highly similar. For example, men and women likely have very similar knowledge of gender stereotypes that associate male with some traits (e.g., strength) and female with other traits (e.g., warmth). Nevertheless, at the center of these networks of broadly shared social knowledge there is a core of social knowledge that differs importantly across society's demographic subgroups. This is the collection of associative links to *self*. The chief theoretical device used in this article has been to spell out the social-cognitive consequences of individual and group differences in self's associative connections.

Predictions 1 and 2 followed from applying the theory's first principle (balance-congruity) to configurations involving varied associations of self to ingroups. Additional predictions can be generated by applying the theory's second and third principles to structures that vary in self's associations. Principle 2 (imbalance-dissonance) describes consequences of self being associated with one of two bipolar-opposed concepts, and Principle 3 (differentiation) describes consequences of self being associated with both of a pair of bipolar-opposed concepts. The method requirements of testing predictions that follow from Principles 2 and 3 are challenging enough so that no definitive tests have yet been conducted. Nevertheless, it is useful to state two further predictions in order both to extend the theory's empirical implications and to provide targets for future tests.

> **Prediction 3: Contrasted identity, self-concept, and attitude**. *Social knowledge structures resist forming associations of ingroup or self to concepts associated with a group that is bipolar-opposed to one's ingroup.*

Prediction 3 follows from the balance-congruity and imbalance-dissonance principles in combination. If self becomes associated with a concept that is linked to a group bipolar-opposed to an ingroup (i.e., to an outgroup), then the balance-congruity would call for development of a link of self to the to the outgroup. The imbalance-dissonance principle postulates resistance to such configurations.

Figure 11 diagrams the consequences of applying the imbalance-dissonance principle to a network fragment that contains a concept linked to an identity that is bipolar-opposed to one's ingroup. For Figure 11's illustration, the bipolar-opposed identities are *male* and *female* and the concept

machinery is (gender-stereotypically) associated with *male* while *self* is associated with *female*. The imbalance-dissonance principle asserts resistance to a node becoming linked to both of a pair of bipolar-opposed nodes. In terms of Figure 11, someone who is female-identified should resist forming an association between *female* and the male-identified concept of *machinery*.

The methods used to test Predictions 1 and 2 did not include any measure of the resistance to association formation that is described in Prediction 3. Although no such test has yet been attempted, a possible method for testing Prediction 3 might be to (a) establish an association of a novel concept to a group bipolar-opposed to an ingroup, and then (b) measure the ease of acquiring an association between ingroup and the novel concept.

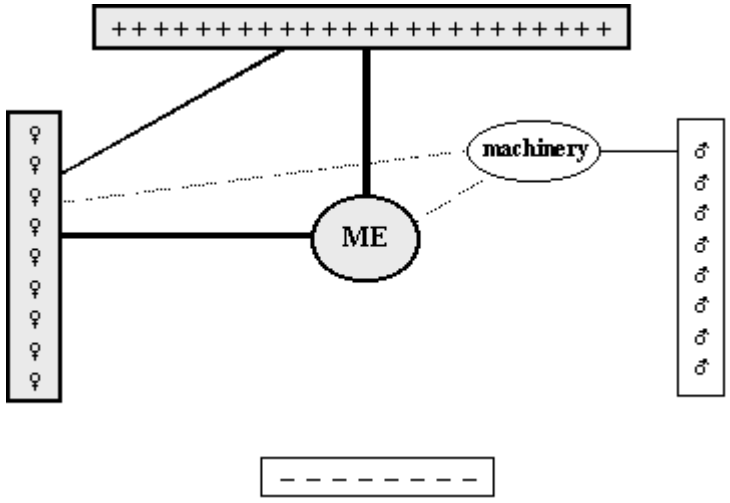

Figure 11. **Resistance to incorporating an outgroup stereotype into self-concept**. The solid links represent associations in a woman's (or girl's) knowledge structure. These include a gender-stereotypic link between the concepts *male* and *machinery*. The dotted links are self-concept (*Me-machinery*) and gender-trait (*machinery-female*) associations that, in accordance with the imbalance-dissonance principle, should resist formation in a knowledge structure that contains the *Me-female* link.

Differentiation of a Group Linked to Bipolar-Opposed Concepts

The imbalance-dissonance principle describes resistance to forming configurations in which a concept is linked to both of two bipolar-opposed nodes. This principle notwithstanding, such 'pressured' concepts (see Definition 3) are of interest because they arise naturally, such as (a) when the social environment changes (for example, if one becomes a member of a previous outgroup), (b) when the social environment is perverse (e.g., if a parent or significant other is alternately loving and punitive), and (c) much more commonly, when some of one's best friends are ____ (fill in the blank with the name of a disliked, stigmatized, or low-status group). Figure 12 diagrams a generalized situation of this last type — a link between self and a member of a group that is associated with negative valence. The effects of this situation, as expected on the basis of Principle 3 (differentiation) are stated as Prediction 4.

Prediction 4: Outgroup differentiation. *Association of self with a member of a disliked outgroup induces differentiation of the outgroup into negatively and positively valenced subconcepts.*

Similar to the situation for Prediction 3, testing Prediction 4 requires methods that are not yet sufficiently developed to provide a test.

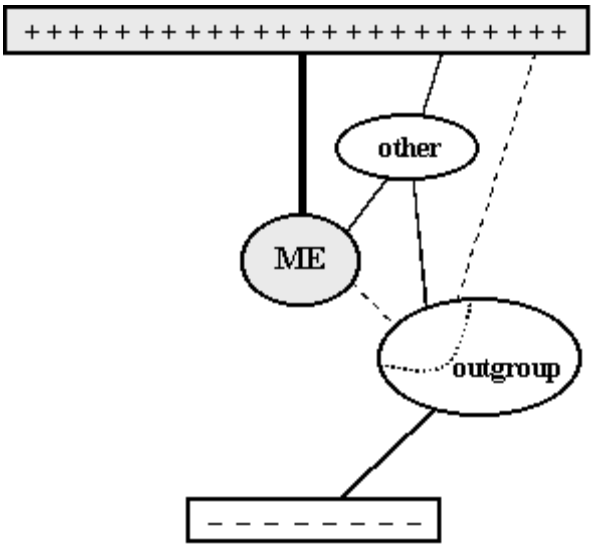

Figure 12. **Differentiation of a pressured-concept configuration.** In this diagram, self (*Me*) is linked to a liked *other* who is a member of (i.e., linked to) a negatively valenced *outgroup*. In this configuration, *Me*, *other*, and *outgroup* are all pressured (in the sense of Definition 3) by virtue of their simultaneous links (direct or mediated) to the two bipolar-opposed valence concepts. This pressure might disappear by differentiation (Principle 3) of any of these three concepts. Differentiation of *outgroup* may the most stable such resolution because this differentiation may be the least likely of the three to be opposed by balance-congruity influences arising outside this fragment. Differentiation, which is indicated by the dotted line that divides *outgroup* into two nodes, permits the formation of the dashed links (expected by the balance-congruity principle) and allows them to exist in a non-pressured configuration.

General Discussion

Relation to Social Identity and Self-Categorization Theories

The constructs of attitude, stereotype, self-concept, and self-esteem are very popular among social psychologists. At least one of these concepts was mentioned in 90% of the 601 articles from the 1996-1998 volumes of *Journal of Personality and Social Psychology* that were accessible through the American Psychological Association's full-text database. An indication of the usefulness of theory that interrelates the four constructs is suggested by observing that two or more of the four constructs were mentioned in 72% of the 601 articles, three or more in 40%, and all four were mentioned in 12% of the articles.

The previous work that most closely shares the present aim of interrelating the four constructs of attitude, stereotype, self-concept, and self-esteem is the well-established body of research and theory on social identity — especially Tajfel's *social identity theory* (Tajfel, 1982; Tajfel & Turner, 1986) and Turner's more recent *self-categorization theory* (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987). Social identity theory (SIT) focuses on intergroup conflict and discrimination associated with group identification. Self-categorization theory (SCT) incorporates and extends social identity theory to a larger collection of social phenomena by placing a social-cognitive account

of the self at its theoretical center (see Turner et al., 1987, pp. 42-43). As with the present unified theory (UT), both SIT and SCT assume a close relation between group membership and self-esteem. All three theories easily generate some similar expectations involving self-esteem, ingroup identity, and ingroup preference. In particular, all three theories expect self-esteem to be enhanced by membership in a valued group and all expect that persons who have strong identification with a membership group should display more positive attitudes toward that group than should those having weak identification.

There are some readily noticeable structural differences between SCT and UT. SCT takes *self-categorizations* ('cognitive groupings of onself and some class of stimuli' [Turner et al., 1987, p. 44]) as its representational building blocks, whereas UT's representational elements are associations. Also, SCT conceives the self as a hierarchical structure of self-categorizations at three levels of abstraction (Turner et al., 1987, p. 45), in contrast to UT's nonhierarchical associative structure. Although these abstract structural aspects of the two theories are sharply different, it is not obvious either that they are theoretically fundamental (for example, the cognitive groupings of SCT might be translatable into the association format of UT) or that they translate to differences in empirical expectations.

A more substantive difference exists between SIT and UT, in their treatments of self-esteem in relation to strength of identification with a novel membership group. This difference can be seen in the two theories' accounts of the role of self-esteem in the *minimal group* phenomenon (Tajfel, Billig, Bundy, & Flament, 1971). The 'minimal group' label comes from the observation that the slightest of bases for establishing a membership relationship to a social group — even to an unfamiliar group — results in the new group member making judgments that are likely to be biased in favor of the group. The present theory's analysis of the relation of self-esteem to ingroup identification can be generalized from Figure 5, replacing the concept *female* in Figure 5 with the novel group. The UT expectation is that greater self-esteem should be associated with greater liking for the new group. (The balance-congruity principle calls for strength of the link between the new group and positive valence to be affected positively by strength of the link of self to positive valence.) In contrast to UT's treatment of self-esteem as an associative connection of self to positive valence, SIT treats self-esteem as a motivational force (desire for a positive self-view) that leads to using group identities as pedestals for downward comparison (i.e., generating positive self-regard by promoting ingroups and/or denigrating outgroups). Whereas SIT therefore predicts bias in favor of a novel membership group to be greater for those who have low self-esteem (Hogg & Mullin, 1999; Rubin & Hewstone, 1998), the UT prediction is that bias in favor of a novel membership group should be greater for those who have high self-esteem.

Other differences of UT from both SIT and SCT stem from the unified theory's consistency principles, which yield the predicted data patterns tested in this article's Experiments 1-5. These are differences in level of detail predicted, rather than mutually opposed predictions.

Perhaps the greatest difference between the unified theory, on the one hand, and SIT and SCT on the other, is an incidental consequence of differences in research methods that have been used in testing the theories. The research programs of SIT and SCT preceded any widespread recognition of the distinction between implicit and explicit measures. Research on SIT and SCT has therefore progressed largely with explicit measures. Tests of UT have been conducted in parallel with implicit and explicit measures, leading to the (so far) consistent finding that patterns predicted by UT are more apparent on implicit than explicit measures. Consequently, it seems likely that tests of UT will proceed primarily with implicit measures, whereas SIT and SCT may remain empirically identified with explicit measures.

Connections to Past, Present, Future

Similarities to past consistency theories. In introducing the three principles of the unified theory, ties of these principles to prior consistency theories were emphasized. The present theory's three principles are especially similar to Heider's (1958) balance theory, moreso than to Festinger's (1957) dissonance theory or to Osgood and Tannenbaum's (1955) congruity theory. Nevertheless, all three of those theories share the central insight that relationships among concepts should tend toward structures organized by a principle of cognitive consistency. Perhaps lost sight of in prior competitions among these theories were (a) that this underlying consistency principle was indeed shared by all of the theories, and (b) that this shared principle has never been seriously questioned either by research data or by abstract theory. This same consistency principle is the central feature of the present theory. The unified theory's balance-congruity, imbalance-dissonance, and differentiation principles state the central consistency theme in different ways, and may eventually be shown to be derivable from a single more general principle (Shoda, Tiernan, & Greenwald, in preparation).

The present theory goes noticeably beyond previous consistency theories in only two respects. First, by virtue of borrowing the format of contemporary neural network representations, the present theory's schematic mental structure (see Figure 1) is built from just two types of entities: *nodes* that represent concepts and *links* that represent associations.[14] Second, the present theory takes advantage of the recent availability of implicit measures for empirical tests. The results of the experiments presented in support of Principle 1 of the theory have so far indicated that the theory's predictions are well realized in implicit (but not explicit) measures.

The present theory has been stated in a way that considers balance as a property of associations at the micro level of triads such as those illustrated in Figures 5, 11, and 12. Most previous consistency theories have similarly focused on isolated fragments of the total cognitive system. However, they have carried the implication – shared by the present theory – that the analysis has the potential to be extended to larger fragments, perhaps even to the entire network (cf. Cartwright & Harary, 1956).

Contemporary influences. Although this article seeks to revive a past era's thinking about cognitive consistency, the devices it uses to do this are contemporary. The article's implicit-explicit distinction draws on a distinction between automatic and deliberate processes that has been made frequently in recent social psychological theory (see Chaiken & Trope, 1999; Wilson, Lindsey, & Schooler, 2000). The implicit-explicit distinction can also be connected to a contrast that academic psychologists studiously avoided through most of the 20th century – the conception of functionally distinct conscious and unconscious modes of cognition. The present theory's associationist account of cognition takes obvious inspiration from connectionist (neural network) theorization that has recently established powerful roots in cognitive psychology, artificial intelligence, and neuroscience. Interestingly, in the late 1950s, when consistency theories dominated social psychology's theoretical landscape, this assocationist approach might have appeared to be an outmoded relic of early 20th century behaviorist theories such as that of Edward Lee Thorndike (who labeled his own approach 'connectionist'; Thorndike, 1932).

---

[14]By contrast, Heider's balance theory distinguished two types of nodes (persons and concepts) and two types of links (unit and sentiment). Dissonance and congruity theories did not use associative network representations. The unified theory's node-and-link format makes it possible to describe attitude, stereotype, self-concept, and self-esteem within a single theoretical framework. It also overcomes some problems that Heider encountered in trying to deal with the concept of a 'not-unit' relation (see Footnote 3).

Possibilities for further unification. The unification that has been achieved in this article is obviously limited. Although the theory was successful in providing an integrated treatment of four important social-cognitive constructs, the data showed that this success was limited to the domain of implicit measures. In addition to trying to accommodate explicit measures, future developments of the present theory might seek to bring descriptions of situational manipulations (such as success and failure) into the same framework with cognitive constructs, and also to extend the theory to cognitions associated with dyadic relations, such as social comparisons among the members of a single group.

Conclusion

This article set out to develop a theoretical integration of social psychology's most important cognitive constructs (stereotype and self-concept) with its most important affective constructs (attitude and self-esteem). Both the effort at unification and the ultimate form of the unified theory were shaped by three influences: (a) growing interest in automatic or implicit social cognition, (b) development of the Implicit Association Test, and (c) formulation of the balanced identity design. The three principles at the core of the unified theory all have roots in social psychology's cognitive consistency theories of the 1950s — especially Heider's (1958) balance theory. In an era that increasingly values recycling of resources, it is a satisfying outcome to reuse the wisdom contained in this classic body of theory.

References

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. (Eds). (1968). *Theories of cognitive consistency: A sourcebook*. Chicago: Rand-McNally.

Banaji, M. R., Nosek, B. A., Greenwald, A. G., & Rosier, M. (1999). Manuscript in preparation.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer (Ed.), *Advances in social cognition*. (Vol. 10, pp. 1-61). Mahwah, NJ: Erlbaum.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (vol. 6, pp. 1-62). New York: Academic Press.

Blair, I., & Ma, J., & Lenton, A. (in press). Imagining stereotypes away: The moderation of automatic stereotypes through mental imagery. *Journal of Personality and Social Psychology*.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of self-esteem: The blind men and the elephant revisted? *Journal of Personality and Social Psychology, 79*, 000-000.

Brunel, F. F., Collins, C. M., Greenwald, A. G., & Tietje, B. C. (1999, October). Making the private public, accessing the inaccessible: Marketing applications of the Implicit Association Test. Paper presented at meetings of the Association for Consumer Research, Columbus, Ohio.

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review, 63*, 277-293.

Chaiken, S., & Trope, Y. (Eds.) (1999). *Dual-process theories in social psychology*. New York: Guilford.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (in press). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*.

Dasgupta, N., & Greenwald, A. G. (in press). Exposure to admired group members reduces automatic intergroup bias. *Journal of Personality and Social Psychology*.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (in press). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*.

Deaux, K., Winton, W., Crowley, M., & Lewis, L. L. (1985). Levels of categorization and content of gender stereotypes. *SocialCognition*; *3*, 145-167.

Farnham, S. D. (1999). *From Implicit Self-Esteem to In-group Favoritism*. Unpublished doctoral dissertation, University of Washington.

Farnham, S. D., & Greenwald, A. G. (1999, June). In-group favoritism = implicit self-esteem × in-group identification. Paper presented at meetings of the American Psychological Society, Denver, CO.

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

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

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Social Psychology, 58*, 203-211.

Greenwald, A. G. (1981). Self and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 15, pp. 201-236). New York: Academic Press.

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

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79*, 1022-1038.

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

Greenwald, A. G., & Pratkanis, A. R. (1984). The self. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 129-178). Hillsdale, NJ: Erlbaum.

Haines, E. L. (1999). *Elements of a Social Power Schema: Gender Standpoint, Self-Concept, and Experience*. Unpublished doctoral dissertation, City University of New York.

Harmon-Jones, E., Mills, J. S. (Eds). (1999). *Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington, DC, USA: American Psychological Association.

Hebb, D. O. (1949). *Organization of behavior*. New York: Wiley.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology, 21*, 107-112.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley.

Hewstone, M., Macrae, C. N., Griffiths, R., & Milne, A. B. (1994). Cognitive models of stereotype change: 5. Measurement, development, and consequences of stereotyping. *Journal of Experimental Social Psychology, 30*, 505-526.

Hogg, M. A., & Mullin, B.-A. (1999). Joining groups to reduce uncertainty: Subjective uncertainty reduction adn group identification. In D. Abrams & M. Hogg (Eds.), *Social identity and social cognition* (pp. 249-279). Oxford, UK: Blackwell.

Hummert, M. L., Garstka, T.A., O'Brien, L., Savundranayagam, M., & Zhang, Y. B. (2000). *Implicit Associations, Age Stereotypes, and Identity*. Manuscript in preparation.

Jacoby, L. L., Lindsay, D. S., & Toth, J. P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist, 47*, 802-809.

Jones, E. E., & Davis, K E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219-266). New York: Academic Press.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15*, 192-238.

Kihlstrom, J. F., & Cantor, N. (1984). Mental representations of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 17, pp. 2-48). New York: Academic Press.

Kihlstrom, J. F., & Klein, S. B. (1994). The self as a knowledge structure. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., vol. 1, pp. 153-208). Hillsdale, NJ: Erlbaum.

Koffka, K. (1935). *Principles of gestalt psychology*. New York: Harcourt, Brace. & World.

Maison, D., Greenwald, A. G., & Bruin, R. (2001 [in press]). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin, 2*, xxx-xxx.

Mellott, D. S., Cunningham, W. A., Rudman, L. A., Banaji, M. R., & Greenwald, A. G. (in preparation). Do the IAT and priming measure the same construct? Evidence for the convergence of implicit measures. University of Washington.

Mellott, D. S., & Greenwald, A. G. (2000). *But I don't feel old! Implicit self-esteem, age identity and ageism in the elderly.* Unpublished manuscript, University of Washington, Seattle, WA.

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

Nosek, B., Banaji, M. R., & Greenwald, A. G. (in press). Harvesting implicit group attitudes and beliefs from a demonstration website. Manuscript submitted for publication.

Nosek, B., Banaji, M. R., & Greenwald, A. G. (2000). Math = Male, Me = Female, therefore Math ≠ Me. Manuscript submitted for publication.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62*, 42-55.

Ottaway, S. A., Hayden, D. C. & Oakes, M. A. (in press). Implicit attitudes and racism: effect of word familiarity and frequency on the Implicit Association Test. *Social Cognition.*

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Gatenby, J. C., Funayama, E. S., Gore, J. C., & Banaji, M. R. (in press). Amygdala activation predicts performance on indirect measures of racial bias. *Journal of Cognitive Neuroscience, 12*, 729-738.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rosenberg, M. J. (1969). The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 279-349). New York: Academic Press.

Rubin, M., & Hewstone, M. (1998). Social identity theory's self-esteem hypothesis: A review and some suggestions for clarification. *Personality and Social Psychology Review, 2*, 40-62.

Rudman, L. A., Ashmore, R. D., & Gary, M. (1999). *Implicit and Explicit Prejudice and Stereotypes: a Continuum Model of Intergroup Orientation Assessment.* Manuscript submitted for publication.

Rudman, L. A., & Glick, P. (in press). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*.

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (in press). Self-esteem and gender identity are manifest in implicit gender stereotypes. *Personality and Social Psychology Bulletin.*

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

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin, 26*, 1315-1328.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing* (2 vols). Cambridge, MA: MIT Press.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501-518.

Shoda, Y., Tiernan, S. L., & Greenwald, A. G. (2000). *An Associative Network Model of Cognitive-affective Consistency Principles*. Manuscript in preparation.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard 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., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99,* 3-21.

Tajfel, H. (Ed.) (1982). *Social identity and intergroup relations.* Cambridge: Cambridge University Press.

Tajfel, H., Billig, M. G., Bundy, R. F., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Psychology, 1,* 149-177.

Tajfel, H., & Turner, J. C. (1985). The social identity theory of intergroup behaviour. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7-24). Chicago: Nelson-Hall.

Tedeschi, J. T., Schlenker, B. R., & Bonomo, T. V. (1971). Cognitive dissonance: Private ratiocination or public spectacle? *American Psychologist, 26,* 685-695.

Thorndike, E. L. (1932). *The fundamentals of learning.* Teachers College, Columbia University: New York.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group.* Oxford: Blackwell.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Nonreactive measures in the social sciences*. Boston: Houghton Mifflin.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences* (2nd ed.). Boston: Houghton Mifflin.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77,* 273-295.

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

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

APPENDIX

Additional Details of Banaji, Nosek, Greenwald, and Rosier (1999)

Banaji, Nosek, Greenwald, and Rosier (1999) used a balanced identity design to investigate racial identity and attitudes. Their design was similar to the gender-attitude study of Farnham and Greenwald (see Figure 7), except for replacing the gender (male-female) contrast with a race (Black-White) contrast. A more minor difference was that, in the self-esteem and race identity IAT measures, the idiographic self-other contrast of Figure 7 was replaced by a generic self-other contrast in which pronouns were used to represent both self (*I*, *me*, *mine*, and *myself*) and other (*they*, *them*, *theirs*, and *other*). In the *race identity* IAT, the race contrast was represented by just the category labels *Black* and *White*. The *race attitude* IAT measure was an average of three IATs that represented the race contrast in different ways (category labels [*Black* and *White*], face pictures [as in Dasgupta et al., 2000], and racially classifiable first names [as in Greenwald et al., 1998]).

Subjects were 61 undergraduate male and female students at Yale University, 30 African American (Black) and 31 European American (White). This experiment was conducted before the requirements for regression analyses of balanced identity designs had been completely formulated. Because it used explicit measures of ingroup identity and ingroup attitude that were worded relative to own race, there was no common numerical scale for White and Black subjects' scores on this measure. Balanced identity analyses were therefore possible only for the implicit measures.

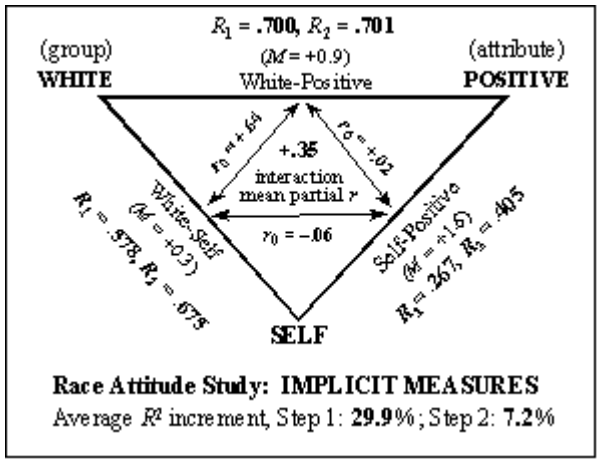

Figure A1. Balanced identity analyses for implicit measures of Banaji et al. (1999). See the caption of Figure 9 for interpretation of the format for presenting these results. The multiple regression with implicit race attitude (White-positive association) as the criterion was entirely consistent with Prediction 1. The regression with implicit self-esteem as the criterion deviated most from Prediction 1, with a greater increment in variance explained at Step 2 than Step 1, and one of the individual-variable coefficients significantly different from zero at Step 2. For both r_0 and R_1, values of .252, .327, .355 are associated, respectively, with p = .05, .01, and .005. (N = 61)

The data for the balanced identity analyses, which are summarized in Figure A1, are considered first descriptively and then in terms of the hierarchical regression tests. The self-esteem measure had the typical strong polarization toward positive values. This led to expecting a positive zero-order correlation between the other two measures, which was observed (r = .64, $p = 10^{-8}$). As expected for a sample that included both races, the mean was near zero for the

measure of implicit race identity (self-White association). The zero-order correlation between the other two measures was therefore also expected to be near zero, and it was (r = .02). The measure of implicit race attitude (White-positive associations) was mildly polarized toward high values, leading to expectation of a weak positive correlation between the other two variables – this was not observed (r = –.06).

All three regression analyses found, consistent with Prediction 1, that the interaction effect term in Step 1 had the expected positive sign. However, for one of the three analyses (the one with implicit self-esteem as the criterion) the amount of variance explained at Step 1 was small (R_1 = .267, p = .04). For all three analyses the interaction effects at Step 2 had the expected positive sign, and two of the three were statistically significant. The three p values for increment in R^2 explained at Step 2 (starting at the top of Figure A1's triangle and proceeding clockwise) were .96, .003, and .05. The latter two significant increments disagreed with expectations from Prediction 1, and there was also a significant coefficient for an individual-variable predictor at Step 2 in the analysis with implicit self-esteem as the criterion. The overall pattern was therefore only partly consistent with Prediction 1. As explained previously, it is possible that the deviations from expectation in Step 2 are due to failures of a scaling assumption rather than to invalidity of theory.

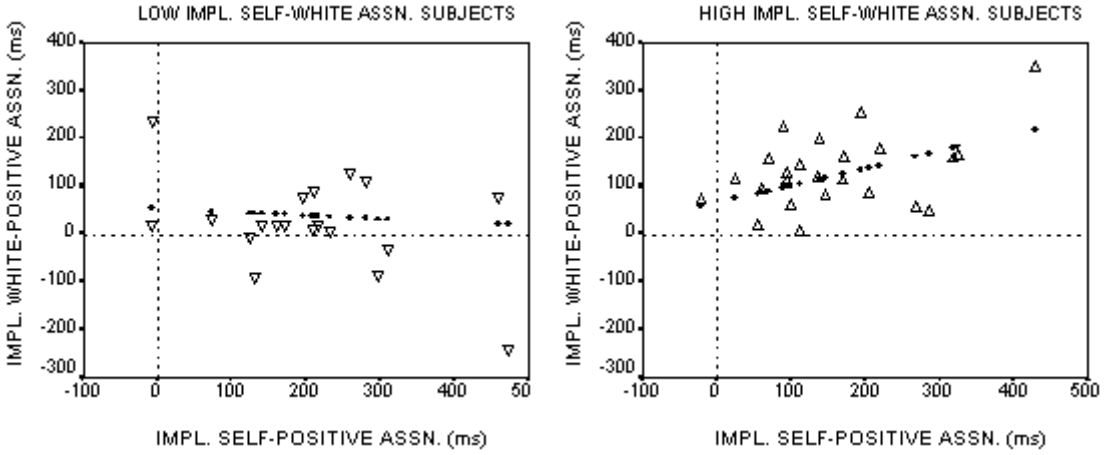

Figure A2. **Example interaction effect for implicit measure data from Banaji et al. (1999)**. The left plot includes data points for all subjects (base-up triangles) whose scores on implicit race identity (self-White association) were at least 0.5 SD units below that measure's mean. The right panel plots those whose scores were at least 0.5 SD units above the mean (base-down triangles). The slopes superimposed on each plot are those expected, from the Step 2 multiple regression results, for the regression of implicit race attitude on implicit self-esteem for hypothetical subjects who have scores on implicit race identity that are one SD below the sample mean (left panel) or one SD above the sample mean (right panel).

To illustrate the balanced identity design's interaction effect test, Figure A2 graphically presents this test from one of the three multiple regressions summarized in Figure A1 – the one with implicit race attitude (White-positive association) as the criterion variable. The left panel of Figure A2 has data for subjects low in implicit White identity (i.e., having implicit Black identity). The expectation for these subjects corresponds to the negatively sloped function in Figure 6 that is labeled 'Low Predictor B'. Similarly, the right panel presents data for subjects high in implicit White identity, corresponding to the positively sloped function in Figure 6 that is labeled 'High Predictor B'. The directions of both slopes corresponded to those predictions, although the slope

in the left panel was only very weakly negative. A significance test for the difference between the slopes in the two panels of Figure A2 is provided by the Step 2 interaction effect, partial $r = .357$, $F(1,57) = 8.34$, $p = .005$.

Additional Details of Mellott and Greenwald (2000)

Mellott and Greenwald (2000) used a balanced identity design to investigate relationships among age identity, ageist attitudes, and self-esteem. The subjects were 52 college students (mean age = 19.7, SD = 1.6) and 46 older subjects (mean age = 74.7, SD = 6.6). All subjects completed both implicit (IAT) and explicit (self-report) measures that (a) compared attitude toward young and old, (b) measured conception of self as young vs. old, and (c) assessed self-esteem (association of self with positive valence).[15] Measures of ageism and age identity were scored such that positive scores indicated preference for old and self-identification as old, respectively.

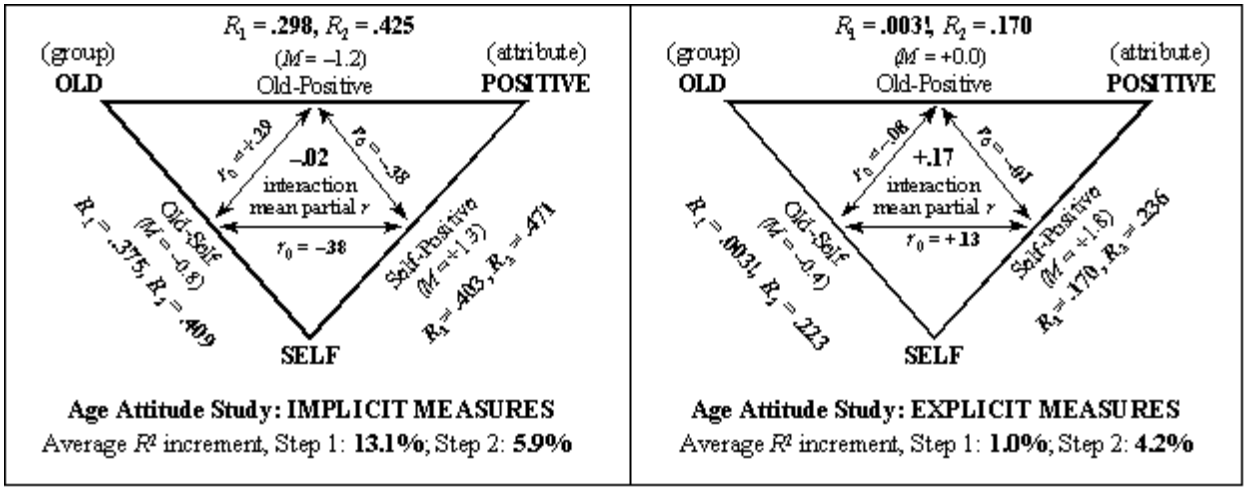

Figure A3. **Summary of balanced identity multiple regression analyses from Mellott and Greenwald (2000).** This data summary again uses the format of Figure 9. The implicit measure data were consistent with the unified theory's Prediction 1 in Step 1 of all three multiple regression analyses, but not in Step 2. The explicit measure data were not at all consistent with Prediction 1. For either r_0 or R_1 in the left panel (N=98), values of .199, .259, and .281 are associated, respectively, with $p = .05$, .01, and .005. For either r_0 or R_1 in the right panel (N=91), values of .206, .269, and .292 are associated with $p = .05$, .01, and .005.

This experiment was conducted with special interest in what it might reveal about implicit cognitions in the sample of older subjects. It was anticipated that the older subjects would show a wide range of attitudes toward the concept of old age, with higher self-esteem elders having a positive attitude toward old age, presumably reflecting psychological comfort with their identity as old. This expectation proved to be approximately the opposite of what the implicit-measure data revealed. That is, the higher the self-esteem of elders the more they both implicitly preferred

[15]Explicit self-esteem was measured with the RSES and explicit attitudes were measured with thermometer and semantic differential scales. Explicit age-identity was measured by two scales. For the first scale, subjects categorized themselves as *very young*, *young*, *middle age*, *elderly*, or *old*. For the second scale, subjects selected the age decade (ranging from *preteen* to *80s*) they felt described them best. The midpoints of these two scales (middle age and age 45, respectively) were treated as zero points indicating equal identification with young and old.

youth to old age and implicitly identified as young rather than old (see zero-order correlations in left panel of Figure A3).

Test of Prediction 1 with implicit measures. Means for the age attitude and age-identity implicit measures were polarized in the negative direction (old associated with negative valence and old associated with other, rather than with self), while the mean value for implicit self-esteem was positively polarized (see means in the left panel of Figure A3). Polarization for the age identity implicit measure was surprising, because inclusion of both young and older subjects was expected to produce a nonpolarized distribution for this measure. With the means actually observed, the one positive and two negative zero-order correlations among the implicit measures were as predicted (and significantly so — see Figure A3). The multiple regression results were also consistent with Prediction 1 at the first step of the analysis. At Step 1, the average standardized b_1 was .359, explaining an average of 13.1% of criterion variance. At Step 2, results for one of the three regressions (the one with implicit age identity as criterion) agreed with Prediction 1. However, inconsistently with Prediction 1, the other two regressions had statistically significant increments in R^2 at Step 2 and had values of either b_2 or b_3 that differed significantly from zero.

Test of Prediction 1 with explicit measures. The three explicit measures showed mean values closer to what was expected for a sample including both young and older subjects. That is, the age identity and age attitude measures were not polarized. The (typical) positive polarization of the explicit self-esteem produced the expectation (from Prediction 1) that the other two measures (age identity and age attitude) would be positively correlated — but they were not (r = –.10). More importantly, the average of the three standardized b_1 values at Step 1 was .055 (none statistically significant, and two negative in sign), explaining only 1.0% of criterion variance and, therefore, providing no support at all for Prediction 1.

The most interesting results of this experiment were the findings that (a) older subjects implicitly identified with young and implicitly preferred young to about the same extent as did young subjects, and (b) these tendencies were strongest among those elders with highest implicit self-esteem. Perhaps these implicit associations of elders can be attributed to their having lived many years in a society that consistently and pervasively values youth over old age. Consistent with the present theory, this should make it psychologically difficult for those with high self-esteem to associate either positive valence or self with old age.[16]

Additional Details of Nosek, Banaji, and Greenwald (2000)

Nosek, Banaji, and Greenwald (2000) investigated self-concepts, gender stereotypes, and attitudes toward academic subject areas in a sample of 46 male and 45 female Yale University undergraduate students. The two academic domains that were contrasted in their association measures were *math* and *arts*. The balanced identity design was expected to reveal that association of *self* with *math* would be consistent with the combination of a gender stereotype that associates *math* with *masculine* and a gender identity as masculine or feminine.

[16]Older subjects in this experiment were also healthy and capable of traveling on their own to the laboratory at which the research was conducted. It would be useful and interesting to have comparable data from a group of less independently functioning elders.

When the Nosek et al. experiment was conducted there was not yet interest in conducting balanced identity designs in parallel for implicit and explicit measures, and its procedure did not include the explicit gender identity measure that was needed for the balanced identity analysis. Accordingly, the balanced identity analysis of Nosek et al. (2000) was limited to implicit measures. For IATs, *mathematics* was represented by 8 items (math, algebra, geometry, calculus, equations, computation, numbers, and Newton) as was *arts* (poetry, art, dance, literature, novel, symphony, drama, and Shakespeare). *Self* and *other* were represented by the usual pronouns. The remaining contrast was *masculine* (brother, father, uncle, grandfather, son, he, his, him) versus *feminine* (sister, mother, aunt, grandmother, daughter, she, hers, her).

Test of Prediction 2. Of the three IATs, the only one that had a polarized distribution was the measure of gender stereotype (male-mathematics association), corresponding to the expected stereotypic association of masculine (more than feminine) with the concept of math. This observation led to the expectation of a positive zero-order correlation between the other two measures, gender identity (masculine-self association) and self-concept (self-math association). That correlation was statistically significant (r = .41, p = .00005). More importantly, results of the 2-step multiple regression analysis were fully consistent with expectations based on Prediction 1. The three standardized values for b_1 in Step 1 were all positive, averaging +.359 (ps $\leq$ .03), and the b_1 coefficients in Step 2 were all positive in sign, averaging +.16 (one was statistically significant). Additionally, the increments in R^2 from Step 1 to Step 2 were all nonsignificant (ps $\geq$ .30) and there were no significant deviations from zero values for the interaction-component predictors in Step 2.

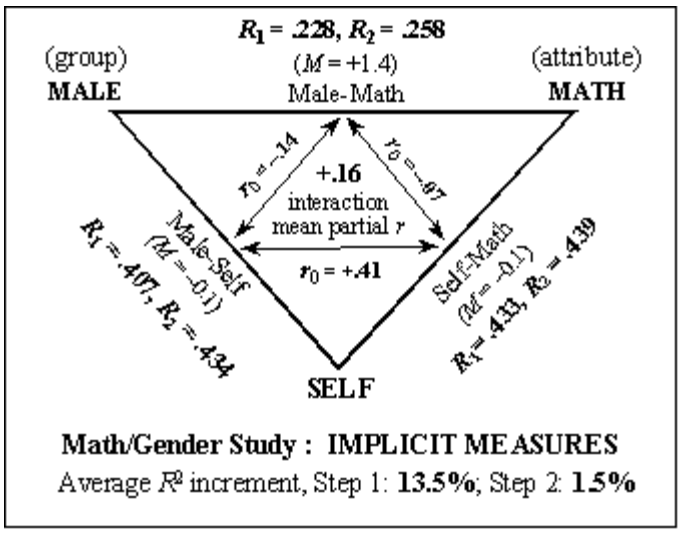

Figure A4. **Summary of implicit measure analyses for Nosek et al. (2000)**. See Figure 9 for explanation of symbols. The implicit measure data were fully consistent with the unified theory's Prediction 1 in both Steps 1 and 2 of all three hierarchical multiple regression analyses. For either r_0 or R_1 (N=91), values of .206, .269, and .292 are associated with p = .05, .01, and .005.