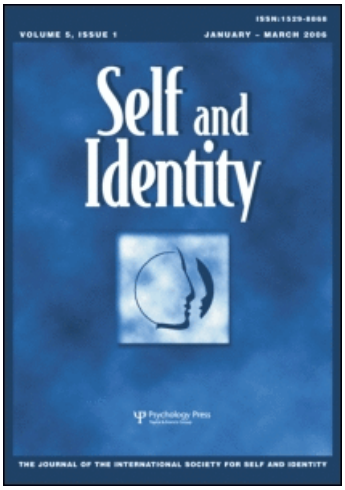

Self and Identity
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713685324

Narcissistic Fragility: Rethinking Its Links to Explicit and Implicit Self-esteem

Aiden P. Gregg [a]; Constantine Sedikides [a]
[a] University of Southampton, Southampton, UK

First published on: 18 April 2009

To cite this Article Gregg, Aiden P. and Sedikides, Constantine(2010) 'Narcissistic Fragility: Rethinking Its Links to Explicit and Implicit Self-esteem', Self and Identity, 9: 2, 142 — 161, First published on: 18 April 2009 (iFirst)
To link to this Article: DOI: 10.1080/15298860902815451
URL: http://dx.doi.org/10.1080/15298860902815451

Self and Identity, 9: 142–161, 2010
http://www.psypress.com/sai
ISSN: 1529-8868 print/1529-8876 online
DOI: 10.1080/15298860902815451

Psychology Press
Taylor & Francis Group

Narcissistic Fragility: Rethinking Its Links to Explicit and Implicit Self-esteem

AIDEN P. GREGG
CONSTANTINE SEDIKIDES

University of Southampton, Southampton, UK

Several studies have tested whether narcissism is a compensatory reaction to underlying ego fragility by examining narcissism's empirical links to both explicit self-esteem (ESE) and implicit self-esteem (ISE), under the general expectation that narcissists should exhibit an abundance of ESE but a dearth of ISE. However, not only have these studies yielded conflicting findings, they have also proceeded from divergent theoretical assumptions that shape the interpretation of their findings. Here, we draw out the implications of three prominent models of the interrelationships between narcissism, ESE, and ISE, before reassessing those interrelationships in a large multi-session study. Two (out of three) indices of ISE covaried negatively with narcissism, consistent with the view that ISE is a global marker for ego fragility. We contextualize our findings in terms of recent research and propose a new mechanism linking ISE to ego fragility.

Keywords: Implicit measures; Implicit self-esteem; Indirect measures; Narcissism; Self-esteem.

Narcissism, once conceptualized as a discrete personality disorder (Akhtar & Thomson, 1982), has been recast as a continuous individual difference (Rhodewalt & Morf, 2005). For research purposes, it is now most often operationalized as high scores on the Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979). Yet the accumulating empirical portrait of the everyday narcissist remains familiar. Such persons are egocentric (Sedikides, Campbell, Reeder, Elliot, & Gregg, 2002), prone to illusions of superiority and specialness (Farwell & Wohlwend-Lloyd, 1998), and liable to be interpersonally abrasive or aggressive (Bushman & Baumeister, 1998).

The above characteristics strongly suggest that narcissists are hypermotivated to self-enhance (Sedikides & Gregg, 2001) or unable to contain their egocentrism (Vazire & Funder, 2006). Most tellingly, narcissists persist with a policy of shameless self-promotion despite the long-term personal and occupational costs of doing so (Morf & Rhodewalt, 2001). Why must narcissists self-aggrandize so relentlessly?

Received 13 December 2007; accepted 26 January 2009; first published online 18 April 2009.

Research reported in this paper was supported by Economic and Social Research Council grant # JW10 R281.

We thank Lisa Baker, Estelle Coe, Tom Cox, Sara Onur, and Carly Rogers for their assistance in data collection.

Correspondence should be addressed to: Aiden P. Gregg, Center for Research on Self and Identity, School of Psychology, University of Southampton, Southampton SO17 1BJ, UK.
E-mail: aiden@soton.ac.uk

If non-narcissists can survive feeling reasonably good about themselves, why do narcissists need to feel so great?

The Nature of Narcissistic Self-regard

Psychodynamic theories, originally developed to explain clinical narcissism, suggest a possible answer: narcissists' excessive efforts to self-promote are symptomatic of deficient rather than overabundant self-regard. In particular, narcissism has been variously put down to compensatory distortions in terms of how one sees oneself and others (Kohut, 1976), attempts to offset insufficient levels of parental love (Kernberg, 1975), or efforts to construct an idyllic false self in place of the imperfect real one (Lowen, 2004). The common thread linking all these theories is that appearance disguises reality: a shaky self hiding behind a puffed-up persona. But is there any concrete evidence for such latent ego fragility in narcissists?

Some findings suggest not. For example, narcissism correlates positively with adaptive traits and negatively with maladaptive traits (Sedikides, Rudich, Gregg, Kumashiro, & Rusbult, 2004), and narcissists are quite certain of who they are (Tschanz & Rhodewalt, 2001). However, other findings subtly implicate ego fragility. For example, narcissists' affective states are more changeable than those of non-narcissists, both in everyday life (Emmons, 1987; Rhodewalt, Madrian, & Cheney, 1998) and in response to experimental manipulations (Bogart, Benotsch, & Pavlovic, 2004). Yet these two sets of findings are not really at odds: the first addresses questions of central tendency, while the second addresses questions of variability. Narcissists can exhibit higher average levels of psychological functioning (i.e., be self-confident high performers overall) while also exhibiting greater fluctuations in psychological functioning (i.e., be dogged by egotistical sensitivity).

Still, it would be useful to have an objective index of the ego fragility hypothesized to underlie narcissism. Psychodynamic hypotheses (e.g., to the effect that X is *really* a sign of not-X) are notoriously tricky to test, and clinical interpretations often lack reliability or validity (Meehl, 1983). Intriguingly, however, the field of social cognition has recently yielded a set of putative indices that may fit the bill.

Implicit Cognition, Measures, and Esteem

Extensive theory and evidence point to the mind operating at two distinct levels (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995), one *explicit* (i.e., controlled, deliberate, logical, conscious, and reflective) and the other *implicit* (i.e., automatic, unintentional, associative, unconscious, and impulsive). Traditional self-report instruments, or *direct measures*, are geared to index the former, whereas some newer cognitive techniques, or *indirect measures* (DeHouwer, 2003), are geared to index the latter. In particular, indirect measures elicit responses under conditions designed to undermine one or more facets of explicit mental processing, such as intentional control (Draine & Greenwald, 1998) or measurement awareness (Fazio, Jackson, Dunton, & Williams, 1995). The upshot is that *implicit associations* between represented constructs can be inferred from patterns of response to stimuli or combinations of stimuli (DeHouwer, 2003). Moreover, given that attitudes can be defined as attribute–object associations (Fazio, 2007), indirect measures can be construed as measures of implicit attitude (Wittenbrink & Schwarz, 2007).

Self-esteem is commonly defined as a positive or negative attitude towards *oneself* (Sedikides & Gregg, 2003). Moreover, the concepts of "positive," "negative," and

"self" can be readily represented—the first two by valenced nouns or adjectives (*heaven* vs. *hell*; *nice* vs. *nasty*), and the third by pronouns, names, or initials (*I* & *me*; *Aiden* & *Gregg*; *A* & *G*). Hence, by inserting such words into indirect measures, it becomes possible to gauge whether and to what extent people implicitly evaluate themselves in a positive or negative way—their level of *implicit self-esteem* (ISE)—on the assumption that self symbols are proxies for the actual self.

The two most commonly used indices of ISE are the *Implicit Association Test* (IAT; Greenwald, McGhee, & Schwartz, 1998), in which respondents attempt to rapidly co-classify stimuli into corresponding categories, and the *Name Letter Task* or *Initials Preference Task* (NLT or IPT; Koole & Pelham, 2003), in which respondents, oblivious to the purpose of the task, rate their liking for letters both in and not in their name. Whichever index is used, *explicit self-esteem* (ESE), measured by conventional questionnaire, turns out to be largely independent of ISE, just as dual-process theories would predict (Bosson, Swann, & Pennebaker, 2000; Rudolph, Schröder-Abé, Schütz, Gregg, & Sedikides, 2008; but see Jordan, Whitfield, & Zeigler-Hill, 2007).

There are also circumstantial grounds for suspecting that ISE might be an objective marker of ego fragility in narcissists. Both pertain to self-evaluation; both reside in the cognitive background; and both can be at odds with what lies in the cognitive foreground. Moreover, if deficits in ISE could be empirically linked to narcissism, this suspicion would be reinforced. As it happens, several studies have already reported such a link (Boldero et al., 2007a; Boucher, 2007; Brown, Bosson, & Swann, 2002; Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003; Rosenthal, 2005; Zeigler-Hill, 2006). However, the nature of the link varies. Moreover, diverging empirical findings are accompanied by differences in theorizing as to why such a link should exist, what form it should take, and whether ESE should be involved. We therefore outline below three competing models of how narcissism relates to both ISE and ESE. We then comment on the viability of these competing models in the light of our own and further findings.

Three Models of the Link between Narcissism and Self-esteem, Explicit and Implicit

The three models are: (1) the *global marker model*; (2) the *full discrepancy model*; and (3) the *partial discrepancy model*. We posit (1). However, other researchers to date have either posited or assumed models (2) and (3). For balance, we articulate all three models side by side.

The global marker model. This model posits that ISE per se is an overall index of ego fragility. In particular, the model posits that ISE is inversely related to ego fragility. Hence, if higher narcissism implies higher ego fragility then higher narcissism should also imply lower ISE. Note that the global marker model makes no reference to ESE. As far as the model is concerned, a positive link between ESE and narcissism exists (at $r \approx .35$; Sedikides et al., 2004) but is irrelevant. Accordingly, the global marker model implies that ESE and ISE should exert separate main effects on narcissism but should not interact (Figure 1, Pattern A). Furthermore, this model does not presuppose any particular theory of what might cause ego fragility in narcissists.

Both indirect and direct empirical findings support the global marker model. First, ISE has been found to covary simply and negatively (albeit sometimes in conjunction with other patterns, or given further conditions) with assorted signs of

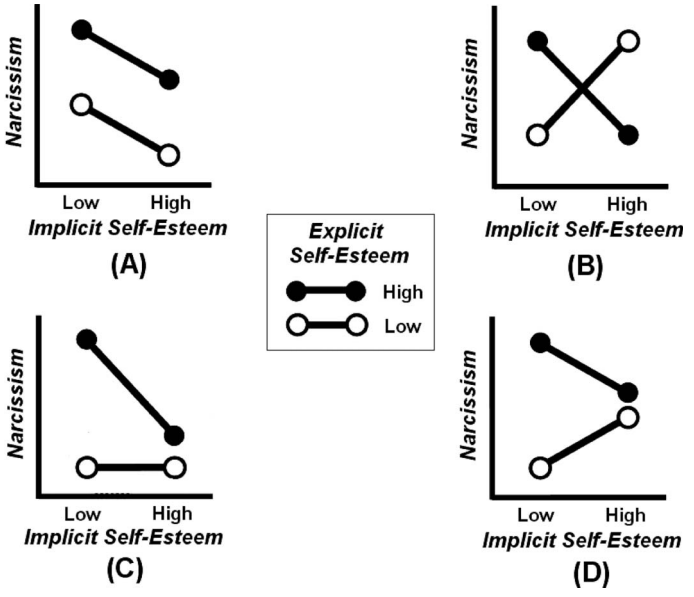

FIGURE 1 chematic depiction of various patterns of predicted mean values derived from hypothetical regressions of narcissism on implicit self-esteem and explicit self-esteem.

ego fragility. Such signs include unstable self-esteem (Zeigler-Hill, 2006), verbal defensiveness (Kernis, Lakey, & Heppner, 2008), reactivity to threat (McGregor & Jordan, 2007), and cognitive reactions to success or failure (Greenwald & Farnham, 2000). Second, ISE has at least twice been found to covary simply and negatively with narcissism (Boldero et al., 2007a; Rosenthal, 2005).

The full discrepancy model. This model posits that ISE and ESE jointly serve as indices of ego fragility in narcissists. Levels of ESE and ISE, each being either high or low, can hence be either congruent or discrepant. According to some researchers, discrepancies between ESE and ISE are either a sign or a cause of ego fragility, and as such are liable to predict or provoke compensatory self-enhancement. For example, Kernis et al. (2005) stated that "fragile self-esteem reflects discrepancies between individuals' explicit and implicit feelings of self-worth. Such discrepancies presumably undermine the certainty and security of individuals' feelings of self-worth, thereby heightening their tendencies to engage in self-protection and self-promotion" (p. 314; see also Jordan et al., 2003, p. 970). Hence, if higher narcissism implies higher ego fragility, then higher narcissism should also imply greater discrepancies between ESE and ISE. Accordingly, the full discrepancy model implies that ESE and ISE should interact to predict narcissism but exert no main effects (Figure 1, Pattern B). Note that the full discrepancy model, unlike the global marker model, postulates a mechanism to generate ego fragility, a potential advantage.

The full discrepancy model has some indirect empirical support, in that discrepant ESE and ISE do interact as specified to predict ego fragility. Schröder-Abé, Rudolph, Wiesner, and Schütz (2007, Study 2) found that people with discrepant ESE and ISE spent less time contemplating rejecting negative feedback than did people with congruent ESE and ISE. Moreover, Briñol, Petty, and Wheeler (2006)

found that discrepancies between implicit and explicit attitudes towards external issues prompted enhanced processing of discrepancy-related information, suggesting by extension that discrepancies between ISE and ESE might be sufficiently disconcerting to prompt attempts at resolution. However, to the authors' knowledge, no direct empirical support yet exists for the full discrepancy model, as far as narcissism per se is concerned. In addition, its underlying rationale is open to question. If discrepancies between ESE and ISE are psychologically disruptive—generating perennial pressure to resolve them (Harmon-Jones & Mills, 1999)—why would trait measures of ESE and ISE remain almost completely uncorrelated? Relatedly, if ESE and ISE are theorized to arise from largely independent psychological systems—which *by definition* do not mutually interfere—then why should ESE–ISE discrepancies be expected to cause psychological disruption in the first place?

The partial discrepancy model. This model, like the full discrepancy model, also posits that ISE and ESE jointly serve as indices of ego fragility in narcissists. However, what distinguishes the partial discrepancy model is the singling out of just *one* cell from the decomposed 2 × 2 interaction of ESE and ISE—typically, the high ESE/low ISE cell—with its value being compared to the values in the other three cells. A higher value in the focal cell is construed as evidence that a combination of higher ESE and lower ISE characterizes people who are higher in narcissism or ego fragility (Figure 1, Pattern C).

Patterns roughly consistent with the partial discrepancy model have emerged empirically. First, various manifestations of ego fragility have been greatest in the high ESE/low ISE cell. These include in-group bias and dissonance reduction (Jordan et al., 2003, Studies 2 and 3), less positive interpretations of ambiguous feedback (Schröder-Abé et al., 2007, Study 1), and increases in conviction strength and estimated consensus following manipulations of uncertainty and failure (McGregor & Marigold, 2003; McGregor, Nail, Marigold, & Kang, 2005). Second, narcissism itself has been found to be highest in the high ESE/low ISE cell (Brown et al., 2002; Jordan et al., 2003; Rosenthal, 2005).

However, the rationale for the partial discrepancy model is also open to question. First, why should the discrepancy between high ESE and low ISE be disruptive but not the discrepancy between low ESE and high ISE? Technically, the magnitude of each discrepancy is equivalent. Second, can the ordinal comparison of one particular cell value to three other cell values do full justice to the *entire pattern* of cell values that emerges? Certainly, it is informative to find that narcissism is often numerically highest when respondents are, say, conjointly higher in ESE and lower in ISE. However, this is *not* equivalent to showing that narcissism overall is significantly higher when both ESE is higher and ISE is lower. The latter is implied *only* by significant main effects of ESE and ISE separately (Figure 1, Pattern A). Moreover, a significant interaction between ESE and ISE, although it *can* be driven solely by an extreme value sufficient to engender collateral main effects of ESE and ISE, can *also* be driven by a variety of other patterns, including one from which a main effect for ISE alone is absent (e.g., Figure 1, Pattern D). But such an interaction would *not* show that narcissism overall is significantly higher when both ESE is higher and ISE is lower. Instead, it would show that narcissism in general is significantly higher when ESE is higher, but that narcissism's link to ISE depends on ESE in a complex way on all four cell values (i.e., that it is more negatively related to narcissism as ESE increases, but more positively related to narcissism as ESE decreases). Still, we do

agree that, whenever ESE × ISE interactions are driven solely by one extreme value in a discrepant cell, their significance can provide a rough means of testing for ego fragility in narcissists.

The Present Research

Given the questions raised by both discrepancy models, we opted to explore afresh the links between narcissism, ISE, and ESE, using the more straightforward global marker model as our theoretical guide. Accordingly, we conducted a sizeable study ($N \approx 200$) in which we predicted that narcissism would covary positively with ESE but negatively with ISE. For methodological completeness, we assessed ISE using three different key measures of ISE. Two were the NLT and the IAT—the current "market leaders." We selected our third measure, the computer-based *Go No-go Association Test* (GNAT; Nosek & Banaji, 2001), given its established viability (Boldero, Rawlings, & Haslam, 2007b), and its conceptual kinship with the IAT. For methodological security, we assessed both computer-based measures of ISE twice to help ensure their reliability, and also administered our measures of ISE, ESE, and narcissism in separate sessions to prevent carry-over effects (Bosson et al., 2000). Finally, we screened all our data carefully, computed our indices of ISE using optimized indices, and checked the psychometric adequacy in advance of our main analyses.

Method

Participants

The sample comprised 206 University of Southampton, UK, undergraduate students. They ranged in age from 17 to 47 ($M = 20.5$, $SD = 4.2$). Most were female (85%), English (94%), and White (94%). They received either payment (£30; approximately $60) or course credit for taking part in the study.

Scheduling, Design, and Procedure

Participants completed six separate sessions. Sessions were run on consecutive days except that at least one week intervened between Sessions 3 and 4. Narcissism was measured in Session 1, ESE in Sessions 2 and 4, ISE (the IAT and GNAT) in Sessions 3 and 5, and ISE again (the NLT) in Session 6.

Participants completed all study sessions on computer, and were urged to avoid distractions. To run each session, participants downloaded an executable program (created using Authorware 5.0; Macromedia, 2000), which when activated initially prompted participants for their password and ID. Upon completing each session, participants immediately returned data files back to the researchers via e-mail attachment, with a success rate exceeding 95%. On completion of the final session, participants were debriefed by e-mail.

Direct Measures

Narcissistic Personality Inventory (NPI). This 40-item questionnaire (Raskin & Hall, 1979) to assess narcissism featured a forced-choice format (Option *A* or *B*).

Sample item: "I am no better or no worse than most people" (A = non-narcissistic) versus "I think I am a special person" (B = narcissistic). Participants responded to each item by clicking one of two labeled onscreen buttons.

The Rosenberg Self-Esteem Scale (RSES). This 10-item questionnaire (Rosenberg, 1965) to assess ESE featured a vertical 4-point scale (from top: *strongly agree*, *agree*, *disagree*, *strongly disagree*). Sample item: "I feel that I have a number of good qualities." Participants responded to each item by clicking one of four labeled onscreen buttons.

Indirect Measures

Two of the three measures of ISE—the IAT and GNAT—were computer-based. In both tasks, participants repeatedly attempted to quickly and accurately classify words into categories. Words changed on each trial, but category labels remained constant across blocks of trials. Average levels of response speed (IAT) or error rate (GNAT) were computed across block, and differences computed to yield indices of ISE. To maximize comparability, the IAT and GNAT featured identical stimulus items (Appendix). These items comprised 12 generically negative words (e.g., *filth*), 12 generically positive words (e.g., *excellent*), 3 other-denoting words (*they*, *them*, *those*), and 3 self-denoting words (*me*, *myself*, [*first name*]—which participants had earlier typed in). Both tasks also featured four identical category labels: *Nasty*, *Nice*, *Not-Me*, and *Me*.

GNAT. We modeled our GNAT on Nosek and Banaji (2001). On each trial, participants had to press, or to refrain from pressing, the *space bar* before a 600 ms deadline elapsed, in order to indicate respectively whether a word did, or did not, belong to two target categories out of a possible four. In the *Me & Nice* block, for example, participants had to press the *space bar* if a word belonged to the categories *Me* or *Nice* (e.g., *myself* or *delight*), but to refrain from pressing it if a word belonged to the categories *Not-Me* or *Nasty* (e.g., *them* or *filth*).

In each block, a pair of target category labels appeared near the top of the screen. On each trial, a word appeared below them for 600 ms (the deadline), regardless of how participants responded. If they responded correctly before the deadline, a green check mark appeared for 100 ms once the word disappeared; if they responded incorrectly, a red X appeared instead. An intertrial interval of 250 ms followed. Key presses made after the deadline were ignored.

Participants' accuracy at distinguishing target and non-target words in each block was assessed. In the *Me & Nice* block, for example, pressing for *myself* or *delight* was coded as a hit, and not pressing as a miss, while not pressing *them* or *filth* was coded as a correct rejection, and pressing as a false alarm. We quantified discriminative accuracy independently of response bias using d' (Green & Swets, 1966). To ensure index computability, a fix-up value of .005 was added to or subtracted from the relevant cell whenever participants' hit or false-alarm rates equaled null or unity, respectively.

In all, the GNAT comprised four separate blocks, presented in random order. Each block featured a different pair of target categories: *Me & Nice*, *Me & Nasty*, *Not-Me & Nice*, and *Not-Me & Nasty*. Each block consisted of 48 randomly ordered experimental trials, immediately preceded by 16 randomly ordered practice trials. Of the 48 experimental trials, 12 featured a *Nice* word, 12 a *Nasty* word, 12 a *Me* word,

and 12 a *Not-Me* word. Each of the 12 *Nice* and 12 *Nice* words was presented once, and each of the three *Me* and *Not-Me* words 4 times. The 16 practice trials contained 4 words of each type. Participants also completed a preparatory GNAT. It comprised two blocks, each containing 24 trials, with *Bird* and *Fish* stimuli replacing *Me* and *Not-Me* stimuli.

We operationalized an implicit preference for self as greater accuracy during compatible blocks (*Me & Nice*; *Not-Me & Nasty*) than during incompatible blocks (*Me & Nasty*; *Not-Me & Nice)*. A composite index of implicit self-esteem ($GNAT_{ALL}$) was derived by: (a) subtracting d' for the *Not-Me & Nice* block from d' for the *Me & Nice* block; (b) subtracting d' for the *Me & Nasty* block from d' for the *Not-Me & Nasty* block; and (c) adding the two resulting difference scores together.

IAT. We modeled our IAT on Greenwald et al. (1998). On each trial, participants pressed a key, on the left or right of the keyboard (*q* or *p*), to classify a word presented in the middle of the screen into one of four categories (*Me*, *Nice*, *Not-Me*, *Nasty*). Each key corresponded to the two categories the labels of which appeared on the corresponding side. After each word appeared, a trial did not advance until participants had pressed a key. Categorization errors were flagged by a red X, adding a 200 ms delay to each trial. Regardless of accuracy, an ITI of 433 ms followed.

The IAT comprised two blocks presented in random order: a *compatible* block in which category labels were configured to reflect a preference for self (*Me* and *Nice* on the upper right, *Not-Me* and *Nasty* on the upper left), and an *incompatible* block in which category labels were configured to reflect a preference for non-self (*Me* and *Nasty* on the upper right, *Not-Me* and *Nice* on the upper left). In other words, the IAT featured only dual-categorization blocks. Previous research indicates the viability of such abbreviations (Teachman, Gregg, & Woody, 2001).

Both the compatible and incompatible block consisted of 48 experimental trials preceded by 4 practice trials. Of the 48 experimental trials, 12 featured a *Nice* word, 12 a *Nasty* word, 12 a *Me* word, and 12 a *Not-Me* word. Each of the 12 *Nice* and 12 *Nasty* words were presented once, and each of the three *Me* and *Not-Me* words 4 times. The 4 practice trials featured one additional word of each type. Trial order, for both experimental and practice trials, was randomized under the constraint that words falling under the *Nice* or *Nasty* categories alternated with those falling under the *Me* or *Not-Me* categories. To mark this distinction, *Nice* and *Nasty* categories and stimuli appeared in blue, *Me* and *Not-Me* ones in black. Participants also completed a preparatory IAT. It comprised both a compatible and an incompatible block, each containing 12 trials, with *Bird* and *Fish* stimuli replacing *Me* and *Not-Me* stimuli.

We operationalized an implicit preference for self as faster completion of the compatible block relative to the incompatible block. To help control for individual differences in response speed, we computed an adjusted difference score (IAT_{ADJ}) that incorporated the essential features of the revised algorithm recommended by Greenwald, Nosek, and Banaji (2003). In particular, we (a) replaced all extreme latencies within each block (<150 ms or >5000 ms) with means derived from the remaining latencies within that block; (b) recoded latencies less than 350 ms or greater than 3000 ms to those boundary values; (c) imposed error penalties by replacing latencies on trials where errors occurred by the mean plus two standard

deviations of the original block latencies; and (d) divided participants' difference scores (i.e., their average latency in the incompatible block minus their average latency in the compatible block) by the standard deviation of original latencies across both blocks.[1]

NLT. The final measure of ISE, administered alone in Session 6, involved a rating task rather than a compatibility task. By clicking one of seven onscreen numerals (1 = *strongly dislike* to 7 = *strongly like*), participants rated their liking for all 26 letters of the English alphabet (plus 19 ASCII symbols, included to defuse suspicion). All characters were capitalized, shown in a font both large (48 point) and familiar (Times New Roman), and presented in random order.

We operationalized an implicit preference for self as a preference for letters in one's name relative to letters not in one's name, controlling for general letter liking. An index reflecting this (NLT_{ALL}) was computed as follows. First, the normative likeability of each letter was estimated by computing the mean liking for it among participants whose names lacked that letter (to ensure no contamination by any implicit preference for self). Second, the rating that each participant gave to each letter was adjusted by subtracting from it the normative likeability of that letter. Third, for each participant, adjusted ratings for letters in their name, and for letters not in their name, were separately averaged, and the latter subtracted from the former. Repeated letters were counted as one.

Results

Missing or Defective Data

Due to participant dropout and technical failure on the one hand, and task non-compliance and extreme scores on the other[2] (see below), 20 participants did not supply usable data for at least one of the first five sessions. Across all sessions, fewer than 3% of the data from explicit measures, and fewer than 6% of the data from implicit measures, were either missing or defective. Listwise *N*s ranged from 185 to 199 across various analyses. A higher dropout rate for the final session led to lower *N*s (114 to 118) for the NLT.

Data Screening Procedures

To protect the integrity of our results, we excluded data that, by conservative criteria, suggested non-compliance (i.e., that strongly implied a pattern of anomalously rapid, slow, mistaken, or stereotyped responding). We either applied criteria recommended by previous researchers (IAT) or criteria we ourselves devised (GNAT, NLT, explicit measures).

GNAT. We considered participants non-compliant on a GNAT if, across all four blocks, they failed both to adequately discriminate targets from distractors and showed a strong bias towards pressing or not pressing. To quantify overall discrimination, we computed the average hit *minus* false-alarm rate across all blocks. On the resulting scale ranging from +24 (perfect target discrimination) to −24 (perfect distractor discrimination), we deemed scores below +5 dubious. To quantify overall bias, we computed the average hit *plus* false-alarm rate across all blocks.

On the resulting scale ranging from 0 (*never pressed*) to 48 (*always pressed*), we deemed scores below 12 or above 36 dubious. We also excluded GNAT data from a session if, in any block, participants responded identically on all trials (i.e., always did or didn't press).

IAT. We considered participants non-compliant on an IAT if, across both blocks, they either made too many errors, or too often responded either too quickly or slowly. In particular, we excluded data if, on either IAT block, we observed an error rate of 20% or higher, or observed extreme latencies (i.e., >150 ms or <5000 ms) on five or more trials.

NLT. On the NLT, we considered participants non-compliant if they gave every letter the same rating.

Explicit measures. Computer administration also enabled us to screen self-reported responses on the basis of their latency (Holden, 1995). On explicit measures, we considered participants non-compliant if six or more of their responses to a questionnaire (e.g., 15% of the NPI) either took less than one second, or more than ten seconds, to complete.

Outliers. Finally, to ensure the robustness of subsequent findings, we screened all explicit and implicit indices for outliers, defined as any value either less than the 25th percentile, or greater then the 75th percentile, by a margin of three times the interquartile range (Tukey, 1977).

Overall Reliability and Positivity Bias

A good index of ISE should possess adequate reliability and validity. We therefore checked our ISE indices for: (a) internal consistency and test–retest reliability; (b) the presence of a strong overall positivity bias; and (c) convergent validity in the form of intercorrelations.

For all indices of ISE, the computation of internal consistency proceeded as follows: data were split into two equivalent sets; corresponding summary values were derived; the correlation between these values was computed; and that correlation was adjusted in line with the Spearman–Brown prophecy formula. For different indices, the equivalent sets differed. In particular, they took the form of: (a) first and last names for NLT_{ALL}; (b) alternate pairs of trials for IAT_{ADJ}; and (c) maximally interspersed (i.e., within randomization constraints) arrays of trials for the $GNAT_{ALL}$. For the IAT_{ADJ} and the $GNAT_{ALL}$, internal consistencies were averaged across Sessions 3 and 5. Test–retest reliabilities took the form of correlations between corresponding indices in those sessions.

All indices of ISE exhibited reasonable levels of internal consistency (Table 1, Column 2). The internal consistency of IAT_{ADJ} was a little lower than usual (Nosek, Greenwald, & Banaji, 2007; $r = .70$ to .90), possibly due to block order being randomized. On the other hand, that of the $GNAT_{ALL}$ was markedly higher than usual (cf. Nosek & Banaji, 2001; $r = .30$), possibly due to the use of practice tasks and trials.

Variation in true scores reduces correlations across time; but so too does measurement error. So, to estimate true underlying change, we disattenuated raw test–retest correlations (Table 1, Columns 3 and 4) using internal consistency

TABLE 1 Explicit and Implicit Indices: Reliabilities and Positivity Bias

Index	Internal consistency	Test–retest (raw)	Test–retest (disattenuated)	One-sample *t*	Cohen's *d*
Explicit					
NPI	.81[a]	–	–	−18.80	−1.34
RSES	.91[b]	.89	.98	12.80	0.91
Implicit					
NLT_{ALL}	.68[d]	–	–	5.04	0.47
IAT_{ADJ}	.60[c]	.31	.51	32.24	2.33
$GNAT_{ALL}$	.75[c]	.51	.68	24.29	1.74

Notes: $N = 185$–200, except for NLT_{ALL} ($N = 118$). For all *t*-values, $p < .0001$. The *t*-values reflect the degree to which scores reliably depart from an index midpoint. The *d*-values quantify the magnitude of these effects. For explicit indices, internal consistencies represent Cronbach's αs. For implicit indices, internal consistencies represent split-half correlations adjusted in line with the Spearman–Brown prophecy formula. Test–retest reliabilities represent correlations between corresponding indices in different sessions, both raw and disattenuated. [a]Computed from Session 1. [b]Averaged across Sessions 2 and 4. [c]Averaged across Sessions 3 and 5. [d]Computed from Session 6.

coefficients. Results suggested that 26% to 46% of the variance in ISE was conserved over one week, depending on the index.

Overall positivity bias (Table 1, Columns 5 and 6) was quantified in terms by how much observed scores departed from their theoretical scale midpoint (e.g., zero for all indices of ISE). Here (and in all subsequent analyses) both RSES scores, and the IAT_{ADJ} and $GNAT_{ALL}$ indices, were averaged across sessions to maximize measurement reliability. All three implicit indices showed a significant and pronounced positivity bias, suggesting that, at least at an aggregate level, they operated as intended. Still, the threshold for positivity varied across indices, with nearly all participants showing a bias on IAT_{ADJ} (99%) and $GNAT_{ALL}$ (96%), but only a majority showing it on the NLT_{ALL} (66%). In addition, although RSES scores lay mostly towards the upper end of the scale, NPI scores lay mostly towards the lower end; that is, most of the participants liked themselves without self-aggrandizing.

Implicit Intercorrelations and Explicit Intercorrelations

Convergence among indices of ISE would be a further indication of validity. Problematically, previous studies have failed to find such convergence (Bosson et al., 2000), a defect that characterizes implicit measures generally (Fazio & Olson, 2003), and is in part due to poor reliability (Cunningham, Preacher, & Banaji, 2001). In our dataset, overall convergence was also modest: all three positive intercorrelations were low but collectively marginal—pooled $N = 143$; $\chi^2(3) = 6.30$, $p = .09$ (Steiger, 1980). Individually, the NLT_{ALL} index did not correlate significantly with either the IAT_{ADJ} index, $r(118) = .11$, or the $GNAT_{ALL}$ index, $r(118) = .10$, both $ps > .10$ (disattenuated $rs = .18$ and .22). However, the IAT_{ADJ} and $GNAT_{ALL}$ indices did correlate significantly, $r(185) = .15$, $p < .05$ (disattenuated $r = .18$). This may have been because these indices were (a) composites designed to assess comparative preference for self versus non-self and (b) derived from measures sharing a similar

modus operandi (see also Rudolph et al., 2008). Moreover, the correlation might have been higher still if block order in both measures had been held constant instead of allowed to vary randomly.

As expected, the NPI and RSES correlated moderately positively, $r = .47$, $p < .0001$.

Explicit–Implicit Intercorrelations

Also as expected (Bosson et al., 2000; Rudolph et al., 2008), none of the three indices of ISE correlated with the RSES (Table 2, upper). However, two of these indices—$GNAT_{ALL}$ and NLT_{ALL}—correlated *negatively* with the NPI—consistent with our predictions, as guided by the global marker model. That is, although ISE and ESE never covaried, ISE and narcissism did, such that higher levels of narcissism implied lower levels of ISE. Moreover, when ESE was partialed out, these two negative correlations persisted, even increasing slightly, both to $r = .21$. The two positive results are consistent with the hypothesis that ISE reflects the fragility of the narcissistic ego, but independently of ESE.

Testing for ESE × ISE Interactions

The foregoing implies that ESE and ISE (two times out of three) exerted main effects on NPI scores. Over and above such main effects, did any interactive effects emerge, as the full and partial models would predict? Following standard procedure (Aiken & West, 1991), we centered our predictors (i.e., the RSES scores and each ISE index), multiplied them to create corresponding interaction scores, and then regressed NPI scores onto the former prior to regressing them onto the latter. In no regression did coefficients for interaction scores even approach significance (Table 3). In other words, no evidence emerged that ESE moderated the simple inverse relation between narcissism and ISE, as indexed by $GNAT_{ALL}$ and NLT_{ALL}.

Supplementary Analysis I: Composite Index of ISE

Did our study reveal a simple inverse link between ISE and narcissism *overall*? To address this question, we standardized and summed our three indices of ISE to create a composite index, ISE_{COMP}, and then re-ran all pertinent analyses. We obtained an affirmative answer: the ISE_{COMP} index was inversely related to narcissism (Table 2, lower) yet did not interact with ESE to predict narcissism (Table 3, lower).

TABLE 2 Correlations between Explicit and Implicit Indices

Index	RSES (raw)	RSES (disattenuated)	NPI (raw)	NPI (disattenuated)
NLT_{ALL}	−.03	−.04	−.20*	−.27
IAT_{ADJ}	.00	.00	−.04	−.06
$GNAT_{ALL}$	−.04	−.05	−.18*	−.23
ISE_{COMP}	−.03	–	−.18*	–

Notes: $N = 185$–188, except for NLT_{ALL} ($N = 115$). *$p < .05$.

TABLE 3 Regressions of Narcissism on Explicit Self-esteem (Rosenberg), Implicit Self-esteem (Various Indices), and their Interaction Scores (Multiplicative Product)

Index	β	t
NLT_{ALL}	−.18	−2.18*
RSES	.52	6.65***
$NLT_{ALL} \times$ RSES	.01	0.12
IAT_{ADJ}	−.03	−0.52
RSES	.48	7.25***
$IAT_{ADJ} \times$ RSES	−.07	−1.17
$GNAT_{ALL}$	−.17	−2.62**
RSES	.46	7.20***
$GNAT_{ALL} \times$ RSES	.06	0.90
ISE_{COMP}	−.16	−2.56**
RSES	.46	7.15***
$ISE_{COMP} \times$ RSES	.01	0.11

Notes: $N = 185$–188, except for NLT_{ALL} and $NLT_{ALL} \times$ RSES ($N = 115$). *$p < .05$; **$p < .01$; ***$p < .0001$.

Supplementary Analysis II: GNAT regressions

The composite $GNAT_{ALL}$ index that inversely predicted narcissism integrated information from all four GNAT blocks. However, the d'-values from different GNAT blocks can also be considered separately, to explore the nuances of implicit self-evaluation. Accordingly, we simultaneously regressed NPI scores onto the d'-values for all four GNAT blocks. Two links emerged: an inverse link to the *Me & Nice* block value ($\beta = -.24$, $p = .01$) and a positive link to the *Not-Me & Nice* block value ($\beta = .18$, $p = .03$). That is, not only did high narcissists harbor *less* positive automatic associations towards *themselves* (i.e., *Me*) but they also harbored *more* positive automatic associations towards everything that was *not* themselves (i.e., *Not-Me*). In contrast, the d'-values for the two blocks incorporating the category *Nasty* were negligible (βs $< .05$, ps $> .64$). These results suggest that, insofar as ISE reflects ego fragility in narcissists, it entails a dearth of implicit liking for self relative to others, not a surfeit of implicit loathing for self relative to others.

Discussion

We found that participants higher in NPI-measured narcissism, despite predictably scoring higher on a standard measure of ESE, nonetheless scored lower on two out of three measures of ISE, and on a composite index of ISE. Hence, our findings are theoretically consistent with the global marker model, but inconsistent with both the full and partial discrepancy models. Nonetheless, previous research has yielded diverse findings. Averaging across different studies, what is the predominant thrust, and what might explain the empirical variation observed? Recently, Bosson et al. (2008) conducted a helpful meta-analysis of several studies examining the relation between narcissism, ESE, and ISE, in cases where the ISE was indexed by the IAT and the IPT. Its results can be fruitfully compared to and contrasted against our own findings.[3]

First, Bosson et al. (2008) found no overall main or interactive effects for the IAT. This matches our own findings. But might the IAT methodology itself have been at fault, rather than the hypothesis that narcissists have fragile egos? Recently, the IAT has been criticized for being susceptible to a *salience asymmetry* confound (Rothermund & Wentura, 2004). It turns out that IAT effects are driven, not only by correspondences in category meaning, but also by correspondences in category salience. In particular, both negative categories (e.g., *Nasty*), and categories involving negation (e.g., *Not-Me*), are liable to occupy the attentional foreground together, thereby facilitating the co-classification of their respective stimuli. Thus, the IAT may be an impure measure of ISE and hence of ego fragility. Nonetheless, the IAT possesses substantial predictive validity in general (Greenwald, Poehlman, Uhlmann, & Banaji, in press), and has predicted manifestations of ego fragility in particular (Greenwald & Farnham, 2000). Complicating matters further, other researchers have recently found that, whereas IAT measures of ISE whose valenced words possess a communal meaning (e.g., *friendly/rude*) do not correlate with narcissism, those whose valenced words possess an agentic meaning (e.g., *assertive/submissive*) correlate *positively* with narcissism (Campbell, Bosson, Coheen, Lakey, & Kernis, 2007). This finding supports the validity of the IAT, but calls into question whether narcissists' egos are fragile overall. Rather, they may even be *less* fragile in domains that matter to them.

Still, our NLT index of ISE did covary negatively with narcissism, consistent with narcissists having more fragile egos. Yet our findings here were at odds with those of Bosson et al. (2008). Their meta-analysis revealed a weak aggregate *positive* link between narcissism and preference for name letters ($r = .09$, $p = .01$), a result suggestive of ego security.[4] However, a methodological issue complicates the matter. Whereas Bosson et al. meta-analyzed people's preferences for their name initials only (i.e., the IPT), we analyzed preferences for *all* their name letters (i.e., the NLT). We did so to capitalize on all available self-evaluative information, and to maximize the probability of obtaining a significant effect. (Indeed, when we computed an IPT index, it failed to correlate significantly with narcissism in our sample: $r = .06$, $p = .60$.) Still, the aggregate positive link found between narcissism and the IPT must be reckoned with. Furthermore, Sakellaropoulo and Baldwin (2007) found that narcissism was higher among people who, after being primed to think narcissistically, rated name letters as less likeable (what they termed communal ISE) but more attractive (what they termed agentic ISE), as opposed to any other combination of these two rating tendencies. This complex three-way interaction merits replication (N was only 40). However, the broad implication—as for the Campbell et al. (2007) findings—is that the narcissistic ego may be selectively secure.

Nonetheless, our GNAT index of ISE *also* covaried negatively with narcissism, thereby providing a conceptual replication of our findings for the NLT. Moreover, a similar inverse link between narcissism and the GNAT also emerged in an unpublished study ($N = 161$) by Boldero et al. (2007a). These researchers employed a two-block version of GNAT, featuring the equivalent of our *Me & Nice* and *Me & Nasty* blocks. In regression analyses, they found that the inverse link with narcissism was driven by respondents' accuracy in the former block ($\beta = -.20$, $p = .02$), but not in the latter ($\beta = -.08$, $p = .36$), suggesting that higher narcissism entails a dearth of automatic positive associations towards self. These findings dovetail with our own. In so doing, they illustrate the potential utility of the GNAT as an index of particular types of automatic associations towards the self, and help to preserve the hypothesis that narcissism represents a dynamic overcompensation for a fragile ego.

Taking a synoptic view of the research literature on the relation of ISE to narcissism, what conclusions might one draw? The aggregate patterns across IAT and IPT indices of ISE do not conform to predictions of the simple marker, full discrepancy, or partial discrepancy models (Bosson et al., 2008). Yet—as the findings for the GNAT index of ISE illustrate—there remain some tantalizing signs that ISE and narcissism may be inversely linked. Still, at the level of individual studies, results are inconsistent and erratic, sometimes matching only one model, sometimes only another. Although some of this variation may reflect random fluctuation, there is no shortage of potential systematic moderators that might cause it. These include, not only the specificity with which ISE is assessed (Campbell et al., 2007; Sakellaropoulo & Baldwin, 2007), but also how narcissism is defined and measured (Bosson & Prewitt-Freilino, 2007) as well as the structural peculiarities of indirect measures (DeHouwer & Moors, 2007) and individual differences in how people respond to them (Cai, Sriram, Greenwald, & McFarland, 2004), not to mention even the gender of participants studied (Tschanz, Morf, & Turner, 1998; and our sample was largely female). Each possibility raises multiple issues beyond the scope of this article. But here we confine ourselves to addressing a more foundational question raised by the conflicting findings to date: why should measures of ISE reflect ego fragility in the first place?

Closing Thoughts

It is commonly claimed that measures of ISE are especially revealing because measures of ESE have drawbacks such as a susceptibility to introspective blind spots (Koole & Pelham, 2003) and self-reporting biases (Jordan et al., 2007). However, it is not sufficient to criticize the validity of the latter to establish the validity of the former. One must be able to give a forthright answer to the following question: *Why* should spontaneous or automatic positive evaluations of self-denoting or self-connoting stimuli reflect the presence of a secure rather than a fragile ego?

We close by outlining a possible mechanism relating ISE to ego fragility. It involves, first, a *causal connection between ESE and ISE*. In particular, and in line with conventional views of automaticity (Wegner & Bargh, 1998), ISE is held to become established over time as multiple distinct episodes of conscious self-evaluation get gradually proceduralized. Self-symbols (e.g., one's name) then take on the automatic valence that attaches to the self. Accordingly, the fewer positive episodes of conscious self-evaluation people enjoy the less positive their ISE will eventually turn out to be.

Second, the mechanism involves the presence of some factor or factors capable of producing a *consistent dearth of positive self-evaluations*. One possibility, of course, is a problematic developmental history, long theorized to underlie narcissism (Kernberg, 1975; Kohut, 1976), and for which some empirical evidence has recently emerged (e.g., Otway & Vignoles, 2006). However, non-family factors also merit consideration, especially given that genes and shared environment are known to be better predictors of the related constructs of self-esteem and self-esteem stability (Neiss, Sedikides, & Stevenson, 2006). In addition, an intriguing genes-by-environment hypothesis suggests itself. If narcissists are dispositionally inclined to self-enhance (Sedikides & Gregg, 2001) they may often trumpet their self-ascribed virtues. However, if they do so before unimpressed audiences, they may equally often receive skeptical feedback. Consequently, despite habitually rehearsing positive self-evaluations in their own minds, narcissists may habitually encounter criticism

in public, thereby disrupting the smooth proceduralization of positive self-evaluations.

Third, the mechanism involves *ISE being more resistant to change that ESE*. It has long been theorized that implicit attitudes are more robust than their explicit counterparts (Wilson, Lindsey, & Schooler, 2000) and some empirical evidence bears out this assertion (Gawronski & Strack, 2004). Admittedly, the matter is complex, because other evidence suggests that implicit attitudes are malleable (Baccus, Baldwin, & Packer, 2004). However, when direct comparisons are made, it appears that, even if implicit attitudes are as easy to induce as explicit ones they subsequently become more difficult to undo (Gregg, Seibt, & Banaji, 2006). Suppose that this is the case. Then, as learning history unfolds, levels of ISE should vary progressively less than levels of ESE. Or, to use a meteorological metaphor, ISE should track the self-evaluative climate whereas ESE should track the self-evaluative weather.

The relative inertia of ISE may help to explain why the cross-sectional correlation between ESE and ISE is low. If ISE changes more slowly than ESE, then there is room for ESE and ISE to diverge over the shorter term, even if they tend to converge over the longer term. For example, suppose someone's ESE was persistently low until adulthood, thereby engendering low ISE. A subsequent rise in adult ESE—perhaps by way of egotistical overcompensation—would then fail to occasion an immediate corresponding rise in ISE. Hence, the two forms of self-esteem would fall out of alignment. Such might be the predicament of the narcissist. Indeed, all three models reviewed represent attempts to statistically unpack this very premise.

Ultimately, the nature of the link between ISE and narcissism remains obscure and contentious, despite some promising findings like those obtained in the current study. However, building and testing putative models of that link, such as the above, may assist in the achievement of clarity and consensus.

Notes

1. Making or not making any of these four adjustments did not affect either the pattern or significance of our results. We report the adjusted data, given that it is designed in principle to reduce the impact of extraneous dispositional variance.
2. Including extreme scores and data suggestive of non-compliance did not affect either the pattern or significance of any results subsequently reported.
3. This meta-analysis included some of the data derived from the research reported here. However, as omitting our data from the meta-analysis leaves the substantive picture unchanged, the comparisons we draw remain valid.
4. Bosson also reported another aggregate finding: a smaller but nonetheless significant interaction between the IPT index and ESE in predicting narcissism ($r = -.06$, $p = .05$). This interaction was driven by participants with lower ISE and lower ESE having conspicuously lower levels of narcissism. However, this pattern was predicted neither by the full discrepancy nor the partial discrepancy model. Hence, regardless of theoretical approach, the results for the IPT index do not support the hypothesis of narcissistic ego fragility.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Akhtar, S., & Thomson, J. A. (1982). Overview: Narcissistic personality disorder. *American Journal of Psychiatry*, *1*, 12–20.

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. *Psychological Science*, *15*, 498–502.

Bogart, L. M., Benotsch, E. G., & Pavlovic, J. D. (2004). Feeling superior but threatened: The relationship of narcissism to social comparison. *Basic and Applied Social Psychology*, *26*, 35–44.

Boldero, J., Hulbert, C., Lim, C. M. J., Wright, B., Aytekin, S., & Meaklim. H. (2007a). *The contributions of personality and self-esteem to overt and covert narcissism*. Paper presented at the Xth Annual Conference of International Society for the Study of Personality Disorders, the Hague, Netherlands.

Boldero, J., Rawlings, D., & Haslam, N. (2007b). Convergence between GNAT-assessed implicit and explicit personality. *European Journal of Personality*, *21*, 341–358.

Bosson, J. K., Lakey, C. E., Campbell, W. K., Zeigler-Hill, V., Jordan, C. H., & Kernis, M. H. (2008). Untangling the links between narcissism and self-esteem: A theoretical and empirical review. *Social and Personality Psychology Compass*, *2*, 1415–1439.

Bosson, J. K., & Prewitt-Freilino, J. L. (2007). Overvalued and ashamed: Considering the roles of self-esteem and self-conscious emotions in covert narcissism. In J. L. Tracy, R. W. Robins, & J. P. Tangney (Eds.), *The self-conscious emotions: Theory and research* (2nd ed., pp. 407–425). New York: Guilford Press.

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

Boucher, H. (2007). Unpublished data, Bates College, Lewiston, ME, USA.

Briñol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. *Journal of Personality and Social Psychology*, *91*, 154–170.

Brown, R. P., Bosson, J. K., & Swann, W. B. (2002). *How do I love me? Self-love, self-loathing, and narcissism*. Unpublished manuscript, University of Oklahoma.

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology*, *75*, 219–229.

Cai, H., Sriram, N., Greenwald, A. G., & McFarland, S. G. (2004). The Implicit Association Test's *D* measure can minimize a cognitive skill confound: Comment on McFarland and Crouch. *Social Cognition*, *22*, 673–684.

Campbell, W. K., Bosson, J. K., Coheen, T. W., Lakey, C. E., & Kernis, M. H. (2007). Do narcissists dislike themselves "deep down inside"? *Psychological Science*, *18*, 227–229.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163–170.

DeHouwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

DeHouwer, J., & Moors, A. (2007). How to define and examine the implicitness of implicit measures. In B. Wittenbrink & N. Schwartz (Eds.), *Implicit measures of attitudes: Procedures and controversies*. New York: Guilford Press.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, *127*, 286–303.

Emmons, R. A. (1987). Narcissism: Theory and measurement. *Journal of Personality and Social Psychology*, *52*, 11–17.

Farwell, L., & Wohlwend-Lloyd, R. (1998). Narcissistic processes: Optimistic expectations, favorable self-evaluations, and self-enhancing attributions. *Journal of Personality*, *66*, 65–83.

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

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

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

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

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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. K. (1998). Measuring individual difference in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E., & Banaji, M. R. (in press). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier made than undone: The asymmetric malleability of automatic preferences. *Journal of Personality and Social Psychology*, *90*, 1–20.

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

Holden, R. R. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioral Science*, *27*, 343–355.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Implicit self-esteem, explicit self-esteem and defensiveness. *Journal of Personality and Social Psychology*, *85*, 969–978.

Jordan, C. H., Whitfield, M., & Zeigler-Hill, V. (2007). Intuition and the correspondence between implicit and explicit self-esteem. *Journal of Personality and Social Psychology*, *93*, 1067–1079.

Kernberg, O. (1975). *Borderline conditions and pathological narcissism*. New York: Jason Aronson.

Kernis, M. H., Abend, T., Shrira, I., Goldman, B. M., Paradise, A., & Hampton, C. (2005). Self-serving responses as a function of discrepancies between implicit and explicit self-esteem. *Self and Identity*, *4*, 311–330.

Kernis, M. H., Lakey, C. E., & Heppner, W. L. (2008). Secure versus fragile high self-esteem as a predictor of verbal defensiveness: Converging findings across three different markers. *Journal of Personality*, *76*, 1–36.

Kohut, H. (1976). *The restoration of the self*. New York: International Universities Press.

Koole, S. L., & Pelham, B. W. (2003). On the nature of implicit self-esteem: The case of the name letter effect. In S. J. Spencer, S. Fein, M. P. Zanna, & J. M. Olsen (Eds.), *Motivated social perception: The Ontario symposium* (Vol. 9, pp. 93–116). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lowen, A. (2004). *Narcissism: Denial of the true self*. New York: Simon & Schuster.

Macromedia. (2000). *Authorware* (Version 5.0) [Computer software]. San Jose, CA: Adobe Macromedia LLC.

McGregor, I. & Jordan, C. H. (2007). The mask of zeal: Low implicit self-esteem, and defensive extremism after self-threat. *Self and Identity*, *6*, 223–237.

McGregor, I., & Marigold, D. C. (2003). Defensive zeal and the uncertain self: What makes you so sure? *Journal of Personality and Social Psychology*, *85*, 838–852.

McGregor, I., Nail, P. R., Marigold, D. C., & Kang, S.-J. (2005). Defensive pride and consensus: Strength in imaginary numbers. *Journal of Personality and Social Psychology*, *89*, 978–996.

Meehl, P. (1983). Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed.), *Testing scientific theories: Minnesota studies in the philosophy of science* (Vol. 10, pp. 349–411). Minneapolis: University of Minnesota Press.

Morf, C., & Rhodewalt, F. (2001). Unraveling the paradoxes of narcissism: A dynamic self-regulatory processing model. *Psychological Inquiry*, *12*, 177–196.

Neiss, M. B., Sedikides, C., & Stevenson, J. (2006). Genetic influences on level and stability of self-esteem. *Self and Identity*, *5*, 247–266.

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

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.

Otway, L. J., & Vignoles, V. L. (2006). Narcissism and childhood recollections: A quantitative test of psychoanalytic predictions. *Personality and Social Psychology Bulletin*, *32*, 104–116.

Raskin, R., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45*, 590.

Rhodewalt, F., Madrian, J. C., & Cheney, S. (1998). Narcissism, self-knowledge organization, and emotional reactivity: The effect of daily experiences on self-esteem and affect. *Personality and Social Psychology Bulletin*, *24*, 75–87.

Rhodewalt, F., & Morf, C. C. (2005). Reflections in troubled waters: Narcissism and interpersonal self-esteem regulation. In A. Tesser, J. Wood, & D. Stapel (Eds.), *On building, defending, and regulating the self* (pp. 127–151). New York: Psychology Press.

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

Rosenthal, S. A. (2005). The fine line between confidence and arrogance: Investigating the relationship of self-esteem to narcissism. *Dissertation Abstracts International*, *66*(05), 2868B (UMI No. 3174022).

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*, 139–165.

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment*, *24,* 273–281.

Sakellaropoulo, M., & Baldwin, M. W. (2007). The hidden sides of self-esteem: Two dimensions of implicit self-esteem and their relation to narcissistic reactions. *Journal of Experimental Social Psychology*, *43*, 995–1001.

Schröder-Abé, M., Rudolph, A., Wiesner, A., & Schütz, A. (2007). Self-esteem discrepancies and defensive reactions to social feedback. *International Journal of Psychology*, *42*, 174–183.

Sedikides, C., Campbell, W. K., Reeder, G., Elliot, A. J., & Gregg, A. P. (2002). Do others bring out the worst in narcissists? The "Others Exist for Me" illusion. In Y. Kashima, M. Foddy, & M. Platow (Eds.), *Self and identity: Personal, social, and symbolic* (pp. 103–123). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sedikides, C., & Gregg, A. P. (2001). Narcissists and feedback: Motivational surfeits and motivational deficits. *Psychological Inquiry*, *12*, 237–239.

Sedikides, C., & Gregg. A. P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.), *Sage handbook of social psychology* (pp. 110–138). London: Sage.

Sedikides, C., Rudich, E. A., Gregg, A. P., Kumashiro, M., & Rusbult, C. (2004). Are normal narcissists psychologically healthy? Self-esteem matters. *Journal of Personality and Social Psychology*, *87*, 400–416.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

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

Tschanz, B. T., Morf, C. C., & Turner, C. W. (1998). Gender differences in the structure of narcissism: A multi-sample analysis of the Narcissistic Personality Inventory. *Sex Roles*, *38*, 863–870.

Tschanz, B. T., & Rhodewalt, F. (2001). Autobiography, reputation, and the self: On the role of evaluative valence and self-consistency of the self-relevant information. *Journal of Experimental Social Psychology*, *37*, 32–48.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vazire, S., & Funder, D. C. (2006). Impulsivity and the self-defeating behavior of narcissists. *Personality and Social Psychology Review*, *10*, 154–165.

Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 446–496). New York: McGraw-Hill.

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

Wittenbrink, B., & Schwarz, N. (Eds.). (2007). *Implicit measures of attitudes*. New York: Guilford Press.

Zeigler-Hill, V. (2006). Discrepancies between implicit and explicit self-esteem: Implications for narcissism and self-esteem instability. *Journal of Personality*, *74*, 119–143.

APPENDIX

Stimuli used in the IAT and GNAT

Me	Not-Me	Nice	Nasty
me	they	excellent	murder
myself	them	heaven	cancer
[first name]	those	joy	war
		trust	disaster
		peace	hatred
		enjoyment	slaughter
		friend	bomb
		honest	agony
		sweetheart	torture
		love	slime
		freedom	filth
		paradise	traitor