Cvencek, D., Meltzoff, A. N., Maddox, C. D., Nosek, B. A., Rudman, L. A., Devos, T. Dunham, Y., Baron, A. S., Steffens, M. C., Lane, K., Horcajo, J., Ashburn-Nardo, L., Quinby, A., Srivastava, S. B., Schmidt, K., Aidman, E., Tang, E., Farnham, S., Mellott, D. S., Banaji, M. R., Greenwald, A. G. (in press). Meta-analytic use of Balanced Identity Theory to validate the Implicit Association Test. *Personality and Social Psychology Bulletin*.

Supplemental Material for

"Meta-Analytic Use of Balanced Identity Theory to Validate the Implicit Association Test"

***Personality and Social Psychology Bulletin* (2020)**
Accepted March, 2020

Supplemental Material Pertaining to the 4-Test Method

Test 1 of the 4-test method examines the regression of a criterion measure of association strength on the multiplicative product of two theoretically specified predictor measures of association strengths. This is done in Step 1 of a 2-step hierarchical regression for each of the three focal measures (SG, SA, and GA) as a criterion. The balance–congruity principle expects the multiplicative product to be a substantial predictor. The entry of a product term on Step 1 of a hierarchical regression differs from the standard procedure for testing product terms (or interaction effects), which is to enter two component variables on Step 1, then enter their product on Step 2. The rationale for the 4-test method's reversal of this standard procedure can be appreciated with a thought experiment using a known pure-multiplicative theoretical model—prediction of the area of a rectangle from length measures of each of two adjacent sides. Using the standard interaction-effect procedure of entering the two length measures separately on Step 1 as predictors of the rectangle's area, each of those two predictors will account for substantial criterion (area measure) variance, leaving relatively little remaining variance to be accounted for when their (theoretically sufficient) multiplicative product is entered on Step 2. This standard test will give no indication that a pure multiplicative model might account for all the predictable criterion variance. In contrast, entry of the multiplicative product on Step 1 will (properly) show that it accounts for 100% of variance, leaving zero variance to be accounted for when the two side-length measures are entered individually on Step 2.

The 4-test method was described further by Greenwald, Rudman, Nosek, and Zayas (2006) in response to a skeptical appraisal of the method provided by Blanton and Jaccard (2006). Considering this past controversy, the zero-point assumption and the regression method are briefly summarized in the main text. The details of the meta-analytical result using the 4-test

method are reported below. The critique by Blanton and Jaccard is considered further in this Supplemental Material, along with new evidence relevant to their critique.

Using the 4-Test Method to Test Pure Multiplicative Models

Main text provided three Equations (SG1, SA1, and GA1) which indicated how each association in a balanced identity study is embedded in an associative network that includes many other associations. Three parallel univariate regressions for Test 1—one for each of the three association measures in a balanced identity design—are described by the Equations SG2, SA2, and GA2. Derived from preceding Equations SG1, SA1, and GA1, these equations include only the measures represented by the variables with filled-triangle subscripts in Equations SG1–GA1. In Equations SG2–GA2, b_1 corresponds to $b_{\blacktriangle}$; b_0 represents summed effects of the unmeasured additional multiplicative predictors in the preceding equations; and "e" combines sources of random error.

$$SG = b_0 + b_1 \cdot SA{\cdot}GA + e \quad \text{(SG2)}$$

$$SA = b_0 + b_1 \cdot SG{\cdot}GA + e \quad \text{(SA2)}$$

$$GA = b_0 + b_1 \cdot SG{\cdot}SA + e \quad \text{(GA2)}$$

Tests 2–4 of the 4-test method are produced by regression Step 2, in which the two association-strength variables that compose the multiplicative predictor on Step 1 are added as individual predictors. If a pure multiplicative model is valid, Step 2 should add zero to the variance explained in Step 1. This is relatively unlikely because of the multiple additional associations that should have impact on the criterion associations in SG2, SA2, and GA2. Appropriately more modest expectations for Step 2 are (a) that the coefficient of the product term will remain positive (Test 2), (b) that the increment in Multiple *R* due to adding the two

individual predictors will be non-significant (Test 3), and (c) that the added two predictors, when tested individually in Step 2, will be non-significant (Test 4).[1]

The appropriateness of the 4-test method was contested (by Blanton & Jaccard, 2006) shortly after it was first proposed. In response, Greenwald, Nosek, and Sriram (2006) used simulations to contrast the 4-test method with Blanton and Jaccard's preferred method, which was the standard simultaneous multiple regression (SMR) significance test for a multiplicative predictor on Step 2 of a hierarchical regression. Greenwald et al. found both that the 4-test method was more sensitive to presence of a pure multiplicative model than was SMR, and that SMR suffered reduced power in detecting pure multiplicative models to the extent that means of the predictor variables deviated from their zero values (a frequent property of real data sets). This disagreement notwithstanding, the material below includes reports of results using Blanton and Jaccard's preferred SMR method, which is provided by Test 2 of the 4-test method.

Effect Size Calculations and Aggregation Methods for 4-Test Method Findings

Test 1. The effect size measure was the coefficient of the product term entered at Step 1, converted to an *r* value. This *r* was obtained separately for the three different types of criterion measures (self–group [SG], group–attribute [GA], and self–attribute [SA]) in each study and was done separately for IAT and self-report measures. In computing weighted averages as aggregate effect sizes for Test 1, each *r* was weighted by its inverse variance ($n - 3$), where n is the number of subjects in each independent sample (Hedges & Olkin, 1985).

[1] With association-strength measures, for the predictions of a positive coefficient for the product term in both Steps 1 and 2 to apply the three measures must be scored so that a combination of three positive scores defines a balanced configuration. For example, because the combination of *self = female*, *self = good*, and *female = good* is balanced, the measures could be scored so that each of those three associations has a numerically positive value. However, the three measures could also be scored so that any two of the three associations had negative scores, allowing four distinct scoring combinations to be used with the 4-test method. (See also Footnote 2 of the main text.)

Test 2. The effect size measure was the coefficient of the product term at Step 2, converted to a signed partial correlation (*pr*). These were weighted and aggregated as in Test 1.

Test 3. The effect size measure was derived from the test of significance of increase in variance explained at Step 2. Each *p* value was converted to a dichotomous indicator (significant versus non-significant at $p = .05$, 2-tailed). Aggregated proportions of significant results could be compared to the chance value of .05 by a binomial test.

Tests 4. The effect size measure was (as in Test 3), derived from the significance of the individual predictors added at Step 2. Each *p* value was converted to a dichotomous indicator of significant versus non-significant, which could be tested by a binomial test.

4-Test Results for IAT and Self-Report Measures

Test 1: Multiplicative product term at Step 1. For IAT measures, the weighted average *r* for the 36 Step 1 standardized regression coefficients (95% confidence intervals in parentheses) were: $r_{SG} = .330$ (±.039), $r_{GA} = .315$ (±.038), and $r_{SA} = .243$ (±.040; see Table S1). For the 16 samples for which tests could be done with self-report measures, weighted averaged effect sizes were: $r_{SG} = .216$ (±.085), $r_{GA} = .201$ (±.120), and $r_{SA} = .190$ (±.039; see Table S2).

For each of the 16 samples for which both IAT and self-report measures were available, a difference score (Z_{diff}) was computed by subtracting Fisher Z-transformed effect sizes obtained with self-report measures from those obtained with IAT measures. Weighted aggregate Z_{diff} scores were tested for difference from zero by random effects tests and were significantly greater than zero for SG measures, $Z_{diff\ SG1} = .159$ (±.108), $p = .004$, GA measures, $Z_{diff\ GA1} = .155$ (±.129), $p = .019$, and SA measures, $Z_{diff\ SA1} = .102$ (±.056), $p = .0004$. These findings show that there was generally stronger evidence for the balance–congruity principle in Test 1 with IAT than with self-report measures.

Test 2: Coefficient of product term at Step 2. The partial regression coefficients in Step 2 were significantly positive for both IAT measures: pr_{SG} = .158 (±.041), pr_{GA} = .137 (±.043), and pr_{SA} = .168 (±.035), and self-report measures: pr_{SG} = .086 (±.047), pr_{GA} = .110 (±.045), and pr_{SA} = .105 (±.047). There were no significant differences between corresponding IAT and self-report partial regression coefficients, *p*s > .82.

Test 3: Significance of increase in criterion variance explained at Step 2. The proportion of significant results ($p \leq .05$, 2-tailed) for Test 3 was greater than the expected Type I error rate of 5% for both IAT (Table S1) and self-report (Table S2) measures. For IAT, the proportions of significant results at Test 3 were 22% (8/36) for tests with SG criterion measures, 25% for GA, and 25% for SA. For self-report measures the corresponding proportions were 50% (8/16) for SG, 38% for GA, and 50% for SA. These percentages were all significantly greater than 5% at $p \leq .0001$, 2-tailed. In sum, Test 3 showed some deviation from a pure multiplicative model for both IAT and self-report measures, and this deviation was substantially greater for self-report than for IAT measures.

Test 4: Statistical significance for individual predictors at Step 2. Consistent with the results for Test 3, results for binomial tests showed proportions of significant findings greater than the null value of .05 for both types of measures, with a higher proportion of significant *p* values for self-report measures. For IAT measure, the proportions of significant results at Tests 4 were 14% (5/36) and 19% for tests with SG criterion measures, 17% for both GA tests, and 17% and 19% for SA. For self-report measures, the corresponding proportions were 25% (4/16) and 31% for SG, 44% and 25% for GA, and 25% for both SA tests.

Passing of all 12 tests. The results of the 4-test method repeatedly found that support for BIT's balance–congruity principle is stronger when tested with IAT measures of association

strengths than when tested with parallel self-report measures. It is easy to interpret the passing of all 12 tests (four for each of the three criterion measures) as support for a pure multiplicative model, which provides strong support for the balance–congruity principle. This was observed, remarkably, 14 times in the 36 samples for IAT measures, and once in 16 samples for self-report measures (see Table S1). The multiple confirmations of pure multiplicative models for IAT measures are remarkable because the associations in each study are embedded in multiple configurations of trios of associations (see Figure 1 in Main Text). The confirmation of a pure multiplicative model therefore suggests something that no study has yet been ambitious enough to test—the possibility that the unmeasured additional trio configurations of Equations SG1, SA1, and GA1 often, themselves, maintain consistency with one another.

Supplemental Material Pertaining to the Within-Study Meta-Analysis

For IAT data, the meta-analytic aggregate of the 36 within-study meta-analyses of Test 1 yielded a weighted average *r* of .285 (MOE = .029, $p < 10^{-16}$). For Test 2, the aggregation produced a weighted average *pr* of .152 (MOE = .035, $p < 10^{-16}$) [2]. For Test 1, 31 (86%) of the individual-study averaged *r* coefficients were significantly positive, and 25 (69%) of the averaged *pr* coefficients for Test 2 were significantly positive. Heterogeneity was non-significant for Test 1 ($Q = 41.8$, $df = 35$, $p = .20$) and only weakly significant for Test 2 ($Q = 52.2$, $df = 35$, $p = .03$).

For self-report data, the aggregation of the 16 within-study meta-analyses of Test 1 yielded a weighted average *r* of .201 (MOE = .065, $p = 3.47 \times 10^{-9}$). For Test 2, the aggregation produced a weighted average *pr* of .104, (MOE = .039, $p = 2.34 \times 10^{-7}$). For Test 1, 11 (69%) of

[2] Here and elsewhere in Results, a *p* value of $p < 10^{-16}$ is reported as an inequality because the meta-analysis program used to compute and test weighted average effect sizes (Lipsey & Wilson, 2001) is limited to displaying a minimum *p* value of 10^{-16}.

the averaged *r* coefficients were significantly positive, as were nine (56%) of the averaged *pr* coefficients for Test 2. Heterogeneity was substantial for Test 1 ($Q = 101.8$, $df = 15$, $p = 10^{-14}$), but only weakly significant for Test 2 ($Q = 27.2$, $df = 15$, $p = .03$).

To compare successes of Tests 1 and 2 for IAT and self-report measures, the within-study aggregate tests were examined just for the 16 samples that had both IAT and self-report measures. For both Tests 1 and 2, each of the 16 samples' difference (IAT minus self-report) in the Fisher Z effect size for the within-study aggregate was computed, and these 16 difference scores were aggregated using random effects models. For Test 1, the weighted average difference was 0.124 (MOE = 0.080, $p = .002$). For Test 2, the weighted average difference was 0.098 (MOE = 0.059, $p = .001$). These difference tests agreed with the previously described results comparing IAT versus self-report effect magnitudes for Tests 1 and 2 done separately for each of the three types of measures (SG, GA, and SA) as criterion.

Supplemental Material Pertaining to Comparing Studies Using Self-Esteem Measures With Those Using Other Self-Concept Measures

As discussed in the main text, the present meta-analysis affords an opportunity to compare evidence from studies involving valence (i.e., self-esteem measures as SA measures) with those involving other attributes (i.e., self-concept measures as SA measures). The available evidence was compared in two ways. First, the *within-study meta-analysis* method was applied to both self-esteem ($k = 22$) and self-concept ($k = 14$) measures. Second, the *aggregated mean* outcomes of Tests 1 and 2 of the 4-test method for these two groups of samples were compared meta-analytically.

Within-Study Meta-Analytic Method

IAT measures. Tests of the weighted aggregate means for Tests 1 and 2 showed that for self-esteem measures, the two weighted aggregate effect sizes were $r = .271$ for Test 1 (MOE = .041, $p < 10^{-16}$) and $r = .105$ for Test 2 (MOE = .038, $p = 10^{-7}$). Heterogeneity was non-significant for Tests 1 and 2 ($Qs \geq 23.8$, dfs = 21, ps ≥ .15). For self-concept measures, the two aggregate effect sizes were $r = .302$ for Test 1 (MOE = .042, $p < 10^{-16}$) and $r = .219$ for Test 2 (MOE = .043, $p < 10^{-16}$). Heterogeneity was non-significant for both Tests 1 and 2 ($Qs \geq 7.91$, dfs = 13, ps ≥ .41). The difference between the weighted aggregate effect sizes of self-esteem and self-concept measures was statistically significant by an independent-samples t-test for Test 2, $t(34) = 2.99$, p =.005, but not for Test 1, $p > .26$. For the self-esteem data, 77% of the averaged Test 1 r coefficients and 50% of the averaged Test 2 pr coefficients were significantly positive. For the self-concept data, 100% of the averaged Test 1 r coefficients and 93% of the averaged Test 2 pr coefficients were significantly positive.

Explicit measures. For studies with self-esteem measures, analyses showed that weighted aggregate effect sizes were $r = .214$ for Test 1 (MOE = .134, $p = .003$) and $r = .033$ for Test 2 (MOE = .118, $p = .58$). Heterogeneity was significant for both Tests 1 and 2 ($Qs > 13.39$, dfs = 4, ps ≥ .009). For self-concept measures, aggregate effect sizes were $r = .201$ for Test 1 (MOE = .026, $p < 10^{-16}$) and $r = .145$ for Test 2 (MOE = .024, $p < 10^{-16}$). Heterogeneity was non-significant for both Tests 1 and 2 (Qs ≥ 7.56, dfs = 10, $ps \geq .43$). The difference between the weighted aggregate effect sizes of self-esteem and self-concept measures was not statistically significant by an independent-samples t-test for either Test 1 or Test 2, ps >. 11. For self-esteem, 60% of the averaged r coefficients for Test 1 and 40% of the averaged Test 2 pr coefficients

were significantly positive. For the self-concept data, 73% of the averaged Test 1 *r* coefficients and 64% of the averaged Test 2 *pr* coefficients were significantly positive.

4-Test Method

As a part of the analyses comparing studies using self-esteem measures with those using other self-concept measures, successes in passing the 4-test method—ranging from 0 to 4 tests passed—were compared.

IAT measures. Using the 4-test method, self-esteem measures ($k = 22$) passed an average of 2.68 (out of 4) tests and self-concept measures ($k = 14$) passed an average of 3.29 tests. This difference was not statistically significant by an independent-samples *t*-test, $t(34) = 1.68$, $p = .102$.

Explicit measures. For success in passing the 4-test method, self-esteem measures ($k = 5$) passed an average of 1.80 (out of 4) and self-concept measures ($k = 11$) passed an average of 2.73 tests. This difference was not statistically significant by an independent-samples *t*-test, $t(14) = 1.66$, $p = .119$.

Table S1

Effect Sizes for Each of the Four Tests of the 4-Test Method for the 36 Independent Samples Providing Implicit Data

Citation	Criterion Association Measure														
	Self-Group					Group-Attribute					Self-Attribute				
	Test 1	Test 2	Test 3	Test 4a	Test 4b	Test 1	Test 2	Test 3	Test 4a	Test 4b	Test 1	Test 2	Test 3	Test 4a	Test 4b
	r	*pr*	*p*	p_{SA}	p_{GA}	*r*	*pr*	*p*	p_{SG}	p_{SA}	*r*	*pr*	*p*	p_{SG}	p_{GA}
Aidman & Carroll (2003)	**.623*****	**.521*****	10^{-7}***	**.571**	10^{-5}***	**.501*****	-.301*	10^{-8}***	10^{-6}***	10^{-4}***	.025	**.230**†	10^{-5}***	**.071**†	10^{-5}***
Ashburn-Nardo (2010)	**.320*****	**.215***	**.293**	**.124**	**.935**	**.371*****	**.130**	**.775**	**.736**	**.485**	**.248****	**.166**†	**.349**	**.185**	**.727**
Banaji et al. (1997)	**.578*****	**.250**†	.003***	**.072**†	**.136**	**.700*****	**.357****	**.964**	**.813**	**.868**	**.267***	**.394*****	.050*	.025*	**.750**
Baron (2003)	.102	**.157**	**.567**	**.315**	**.556**	.122	**.134**	**.837**	**.566**	**.916**	.133	**.155**	**.598**	**.321**	**.969**
Cvencek et al. (2016, Study 1)	**.615*****	**.293**†	**.445**	**.495**	**.207**	**.470*****	**.259**	**.164**	**.152**	**.090**†	**.571*****	**.419****	**.371**	**.816**	**.181**
Cvencek et al. (2016, Study 2)	**.321*****	-.104	.001***	**.074**†	2^{-4}***	**.323*****	-.085	.002***	10^{-4}***	**.416**	-.007	-.002	**.299**	**.123**	**.401**
Cvencek et al. (2016, Study 3)	**.316***	**.052**	**.122**	**.580**	.042*	**.321***	**.058**	**.177**	**.064**†	**.847**	.205	**.170**	**.872**	**.785**	**.757**
Cvencek et al. (2011)	**.214*****	**.151***	**.329**	**.141**	**.912**	**.164***	**.170***	**.569**	**.699**	**.365**	**.200*****	**.163***	**.079**†	**.106**	**.163**
Cvencek et al. (2014)	**.231*****	**.142**†	**.091**†	.029*	**.680**	**.179***	**.165***	**.761**	**.817**	**.462**	**.226*****	**.137**†	**.056**†	.021*	**.407**
Devos, Blanco, Muñoz, et al. (2008)	**.205***	**.093**	**.872**	**.636**	**.961**	**.207***	**.084**	**.302**	**.962**	**.129**	**.201***	**.179***	**.099**†	**.655**	.046*
Devos, Blanco, Rico, et al. (2008)	**.335*****	**.296*****	**.779**	**.633**	**.553**	**.316*****	**.322*****	**.522**	**.664**	**.333**	**.340*****	**.293*****	**.973**	**.880**	**.848**
Devos & Cruz Torres (2007, Study 1)	**.571*****	**.236***	**.769**	**.986**	**.473**	**.555*****	**.197**†	**.441**	**.228**	**.612**	**.273***	**.281***	**.660**	**.659**	**.668**
Devos & Cruz Torres (2007, Study 2)	**.458*****	**.351***	**.299**	**.134**	**.213**	**.511*****	**.310***	**.461**	**.216**	**.368**	**.522*****	**.283**†	**.466**	**.264**	**.584**
Devos et al. (2007, Study 3)	**.299***	**.237**†	**.641**	**.485**	**.403**	**.341****	**.307***	**.309**	**.137**	**.190**	**.303***	**.333***	**.343**	**.178**	**.417**
Devos et al. (2010, Study 2)	**.467*****	**.271****	**.469**	**.227**	**.627**	**.414*****	**.209***	.012*	**.764**	.003***	**.328*****	**.227***	.018*	**.246**	.005***
Dunham et al. (2007)	**.213***	**.170***	**.085**†	**.901**	.029*	.162†	**.162**†	.029*	.028*	**.147**	.130	**.133**	**.566**	**.901**	**.287**
Dunham et al. (2007)	.074	**.017**	**.400**	**.529**	**.221**	.074	**.027**	**.467**	**.242**	**.761**	.032	**.031**	**.804**	**.545**	**.751**
Farnham & Greenwald (1999)	**.472*****	**.269***	**.483**	**.252**	**.607**	**.445*****	**.100**	**.727**	**.863**	**.480**	**.428*****	**.216**†	**-.591**	**.389**	**.894**
Gumble & Carels (2012)	.055	-.038	**.515**	**.768**	**.337**	.191†	-.011	**.192**	**.344**	**.071**†	.150	-.029	**.176**	**.684**	**.098**†
Horcajo et al. (2010, Study 3)	**.509***	-.155	**.169**	**.069**†	**.163**	.400†	-.171	**.251**	**.141**	**.962**	.149	-.412†	.038*	.012*	**.345**

Horcajo et al. (2010, Study 3)	.114	-.285	**.224**	**.538**	**.089**†	.067	-.264	**.218**	**.101**	**.574**	-.279	-.359†	**.493**	**.362**	**.350**
Horcajo et al. (2010, Study 4)	**.397***	-.011	**.524**	**.261**	**.497**	.255	-.283	**.127**	**.063**†	**.124**	.270	-.086	**.257**	**.183**	**.377**
Horcajo et al. (2010, Study 4)	.268	-.054	.007**	.002**	.043*	.166	**.009**	.004***	**.194**	.003***	-.186	**.085**	.001***	.002***	.003***
Lane et al. (2005)	**.227*****	**.032**	**.297**	**.529**	**.120**	**.298*****	**.140***	**.540**	**.273**	**.601**	**.209*****	**.104**	**.829**	**.900**	**.541**
Lane et al. (2005)	**.303*****	**.088**	**.151**	**.909**	**.139**	**.345*****	**.136***	**.345**	**.291**	**.815**	**.266*****	**.103**	**.899**	**.838**	**.833**
Mellott & Greenwald (2000)	**.375*****	**.019**	**.224**	**.089**†	**.317**	**.298*****	-.112	.007**	**.059**†	.003***	**.403*****	**.031**	.031*	**.174**	.011*
Meltzoff et al. (2019)	**.327*****	**.186**†	**.334**	**.308**	**.660**	**.349*****	**.154**	**.722**	**.456**	**.633**	**.236***	**.219***	**.650**	**.355**	**.993**
Meltzoff et al. (2019)	**.349*****	**.180**†	**.949**	**.746**	**.856**	**.244***	**.109**	**.471**	**.508**	**.309**	**.222***	**.161**	**.434**	**.987**	**.200**
Nosek et al. (2002, Study 2)	**.407*****	**.094**	**.295**	**.227**	**.283**	**.228***	**.216***	**.511**	**.344**	**.828**	**.433*****	**.173**	**.741**	**.445**	**.938**
Nosek & Smyth (2011)	**.309*****	**.219*****	.029*	.056*	.016*	**.347*****	**.265*****	10^{-7}***	.005***	10^{-7}***	**.380*****	**.227*****	10^{-6}***	**.153**	10^{-7}***
Rudman et al. (2001, Study 4)	**.363*****	**.390*****	**.088**†	.029*	**.589**	**.355*****	**.336*****	**.996**	**.942**	**.967**	**.296*****	**.343*****	**.094**†	.035*	**.772**
Rudman & McLean (2013, Study 1)	**.442*****	**.178***	**.629**	**.957**	**.352**	**.430*****	**.194****	**.103**	**.302**	**.184**	.134†	**.156***	**.427**	**.849**	**.195**
Schmidt & Nosek (2015)	**.342*****	**.096*****	10^{-17}***	**.197**	10^{-16}***	**.342*****	**.087*****	10^{-17}***	10^{-17}***	10^{-5}***	**.220*****	**.142*****	.003***	**.853**	.001***
Srivastava & Banaji (2011)	.135	**.231***	.016*	.022*	**.052**†	.144	**.276*****	.011*	.003***	**.256**	.152	**.225***	**.080**†	**.025**	**.724**
Steffens et al. (2010, Study 1)	**.399*****	**.305*****	10^{-3}***	10^{-4}***	**.706**	**.305*****	**.230****	**.408**	**.655**	**.289**	**.370*****	**.313*****	10^{-4}***	10^{-4}***	**.316**
Tang & Greenwald (2013)	**.317***	**.045**	**.233**	**.223**	**.181**	**.309***	**.066**	**.234**	**.250**	**.187**	**.331****	**.086**	**.233**	**.339**	**.145**
Average effect size	.330	.158	.315	.364	.360	.315	.137	.352	.358	.397	.243	.168	.317	.399	.436
(95% CI)	(±.039)	(±.041)				(±.038)	(±.043)				(±.040)	(±.035)			
p	10^{-38}	10^{-13}				10^{-38}	10^{-9}				10^{-38}	10^{-38}			
p[Q]	10^{-5}	.0001				.0001	10^{-5}				.0001	.023			

Note. Balanced identity design always includes measures of associations that link the concept of *self* with one *group concept* (e.g., male) and one *attribute concept* (e.g., valence); Effect sizes for Tests 1 and 2 (*r*s) are presented separately for each of the three regressions in which one measure of association strength is always entered as a criterion (e.g., measure of the *self-group* association) and the other two measures as predictors (e.g., measures of *group-attribute* and *self-attribute* associations). Test 1 is always tested at the regression Step 1 and Tests 2–4 are always tested at the regression Step 2. The weighted mean effect sizes at the first regression step (*r*), their 95% confidence intervals (CIs), transformed back to the *r* metric were computed from a random-effects test for Fisher's *Z*-transformed *r* values at Step 1 of a multiple hierarchical regression analysis. Effect sizes for Tests 3 and 4 are reported as average *p* values at Step 2 (see text for details). *pr* = signed, partial correlation coefficient for the product term at Step 2; *p* = *p* values indicating statistical significance of increase in R^2 at Step 2; p_{SG}, p_{GA} and p_{SA} = *p* values indicating statistical significance of individual SG, GA, and SA

predictors added at Step 2; *p* [Q] = probability values for fixed-effects test of homogeneity (Hedges & Olkin, 1985). Bold font indicates passed tests. † = .05 < $p \leq .10$; * = .01 < $p \leq .05$; ** = .005 < $p \leq .01$; *** = $p \leq .005$

Table S2

Effect Sizes for Each of the Four Tests of the 4-Test Method for the 16 Independent Samples Providing Explicit Data

Citation	Criterion Association Measure														
	Self-Group					Group-Attribute					Self-Attribute				
	Test 1	Test 2	Test 3	Test 4a	Test 4b	Test 1	Test 2	Test 3	Test 4a	Test 4b	Test 1	Test 2	Test 3	Test 4a	Test 4b
	r	*pr*	*p*	p_{SA}	p_{GA}	*r*	*pr*	*p*	p_{SG}	p_{SA}	*r*	*pr*	*p*	p_{SG}	p_{GA}
Ashburn-Nardo (2010)	**.337*****	-.060	10^{-5}***	**.560**	10^{-4}***	**.411*****	-.187*	10^{-6}***	10^{-6}***	.002***	.094	-.060	.014*	**.565**	.019*
Cvencek et al. (2011)	**.163***	**.108**	**.065**†	.030*	**.679**	.124†	**.117**†	**.166**	**.450**	**.146**	**.209*****	**.142***	.026*	.022*	**.134**
Cvencek et al. (2014)	**.264*****	**.225*****	**.201**	**.134**	**.271**	**.225*****	.223***	**.347**	**.298**	**.240**	**.268*****	**.239*****	**.195**	**.083**†	**.495**
Devos, Blanco, Rico, et al. (2008)	.120	**.007**	**.539**	**.267**	**.632**	.077	**.035**	**.542**	**.849**	**.280**	.118	**.009**	**.354**	**.319**	**.271**
Devos & Cruz Torres (2007, Study 1)	.161	**.165**	**.161**	**.556**	**.238**	.128	**.189**†	**.393**	**.174**	**.984**	.213†	**.146**	**.514**	**.301**	**.748**
Devos & Cruz Torres (2007, Study 2)	**.526*****	-.081	**.375**	**.195**	**.405**	**.501*****	**.208**	**.266**	**.337**	**.171**	.195	**.125**	**-.424**	**.611**	**.404**
Devos et al. (2010, Study 2)	**.398*****	**.137**	.009**	.006**	**.948**	**.218***	**.170**†	**.890**	**.697**	**.909**	**.344*****	**.143**	.002***	.001***	**.977**
Farnham & Greenwald (1999)	-.132	**.082**	.022*	**.120**	**.105**	-.126	**.168**	**.080**†	**.058**†	**.262**	-.002	**.007**	**.138**	**.200**	**.989**
Mellott & Greenwald (2000)	-.001	**.139**	**.159**	**.314**	**.133**	-.002	**.203**†	**.129**	.044*	**.318**	.172	**.184**†	**.415**	**.197**	**.669**
Meltzoff et al. (2019)	.179†	**.174**†	**.767**	**.733**	**.489**	**.207***	**.225***	**.572**	**.357**	**.670**	**.238***	**.233***	**.708**	**.527**	**.578**
Meltzoff et al. (2019)	.114	-.008	**.271**	**.190**	**.713**	.035	**.051**	**.240**	**.531**	**.188**	.176†	**.024**	**.225**	**.193**	**.187**
Nosek & Smyth (2011)	**.221*****	**.138*****	10^{-13}***	10^{-11}***	10^{-4}***	**.166*****	**.168*****	.001***	10^{-4}***	**.683**	**.240*****	**.150*****	10^{-8}***	10^{-9}***	**.919**
Rudman et al. (2001, Study 4)	-.003	**.244***	10^{-4}***	**.571**	10^{-4}***	-.108	**.094**	.006**	.002***	**.365**	.176†	**.127**	**.686**	**.832**	**.470**
Rudman & McLean (2013, Study 1)	**.232*****	-.008	3^{-12}***	**.648**	10^{-12}***	**.404*****	-.098	10^{-13}***	10^{-12}***	10^{-6}***	**.177***	-.176*	10^{-7}***	**.968**	10^{-8}***
Schmidt & Nosek (2015)	**.426*****	**.043*****	10^{-227}***	10^{-14}***	10^{-191}***	**.486*****	**.119*****	10^{-158}***	10^{-154}***	10^{-4}***	**.166*****	**.170*****	10^{-19}***	10^{-19}***	10^{-10}***
Srivastava & Banaji (2011)	**.245***	**.010**	.046*	**.883**	**.075**†	**.289*****	**.061**	.006**	.041*	.015*	-.010	**.003**	.028*	**.910**	.015*
Average *r*	.216	.086	.164	.325	.293	.201	.110	.227	.240	.327	.190	.105	.180	.358	.430
(95% CI)	(±.085)	(±.047)				(±.120)	(±.045)				(±.039)	(±.047)			

p	10^{-6}	.0001		.001	10^{-6}		10^{-38}	10^{-5}
p[Q]	10^{-38}	.0001		10^{-38}	.002		.017	.001

Note. Balanced identity design always includes measures of associations that link the concept of *self* with one *group concept* (e.g., male) and one *attribute concept* (e.g., valence); Effect sizes for Tests 1 and 2 (*r*s) are presented separately for each of the three regressions in which one measure of association strength is always entered as a criterion (e.g., measure of the *self-group* association) and the other two measures as predictors (e.g., measures of *group-attribute* and *self-attribute* associations). Test 1 is always tested at the regression Step 1 and Tests 2–4 are always tested at the regression Step 2. The weighted mean effect sizes at the first regression step (*r*), their 95% confidence intervals (CIs), transformed back to the *r* metric were computed from a random-effects test for Fisher's *Z*-transformed *r* values at Step 1 of a multiple hierarchical regression analysis. Effect sizes for Tests 3 and 4 are reported as average *p* values at Step 2 (see text for details). *pr* = signed, partial correlation coefficient for the product term at Step 2; *p* = *p* values indicating statistical significance of increase in R^2 at Step 2; p_{SG}, p_{GA} and p_{SA} = *p* values indicating statistical significance of individual SG, GA, and SA predictors added at Step 2; *p* [Q] = probability values for fixed-effects test of homogeneity (Hedges & Olkin, 1985). Bold font indicates passed tests. † = $.05 < p \leq .10$; * = $.01 < p \leq .05$; ** = $.005 < p \leq .01$; *** = $p \leq .005$

Supplemental Findings Pertaining to Validity of the IAT's Zero Point

Using a higher precision test than previously available, the main text reported strong confirmations of validity of the IAT's theoretically specified (rational) zero-point location. Presented here is an additional relevant interpretation of the IAT's zero point.

Blanton et al.'s (2015) Test of the IAT's Zero-Point Interpretation

A method of assessing the validity of the IAT's zero-point was proposed by Blanton, Jaccard, Strauts, Mitchell, and Tetlock (2015). However, their method had problems that rendered it unsuitable for that purpose.

On self-report attitude measures, higher numbers typically indicate greater liking or favorableness toward the attitude's object. For example, the numerically high end of a thermometer-format measure of attitude toward a political candidate indicates maximum warmth (i.e., favorability) toward the candidate while the low end indicates maximum coldness (i.e., unfavorability). If the measure is scored from 0 to 10, the middle value (5) may be labeled "neither warm nor cold." This midpoint can be understood as an appropriate zero-point, dividing responses into favorable (>5) and unfavorable (<5) to the candidate. Similarly, the midpoint on the widely used Rosenberg (1965) self-esteem inventory, achieved by agreeing equally with self-praising and self-critical statements, is assumed to separate those who are attitudinally positive versus negative toward themselves.

One obtains a score of zero on an IAT attitude measure by responding equally rapidly in the IAT's two combined tasks. The IAT differs from the single-object thermometer measure described in the preceding paragraph because it includes two attitude objects. A political IAT might compare Candidate A with Candidate B, with zero presumably separating respondents who have more positivity toward A from those who have more positivity toward B. This zero-

point is comparable to that for a thermometer-difference measure, in which one responds to a thermometer measure separately for each candidate. The thermometer difference combines these two measures into a relative preference, which produces a zero value when the two candidates have equal thermometer scores.

Blanton and Jaccard (2006) proposed that location of the zero point of IAT measures is "arbitrary" and that "the assumption that the zero point on the IAT measure maps directly onto the true neutral preference [e.g.,] for Whites over Blacks is dubious" (p. 34). Blanton and colleagues (2015) went further to say that the zero point of the race attitude IAT should be placed at a numerically positive value of the IAT's *D* measure.[3] They did not offer a psychological explanation for this presumed displacement of the zero point, but they did propose a statistical regression method to test whether the zero point was displaced in this fashion. Using data they selected to examine with their regression method, they found that the race attitude IAT had an average "right shift" (their term) of the race attitude IAT's zero point of about 1.5 standard deviations above the IAT measure's $D = 0$ value. That estimated average correction would decrease the proportion of people estimated as showing more than slight implicit White preference in the studies they reviewed (pp. 1472–1473) from an average of 83% (using an unaltered IAT *D* measure) to an average of 28%.

In Blanton et al.'s (2015) regression test method, race attitude IAT scores were regressed onto other measures that Blanton et al. believed to have (on average) valid zero points. They expected these analyses to reveal "the mean IAT score one expects to observe among individuals

[3] This assertion applied specifically to the Black–White race attitude IAT, for which a positive *D* score indicates preference for White relative to Black. Scoring direction is arbitrary for IAT measures, at the discretion of researchers. Blanton et al. were not assuming that, if this IAT were scored in the reverse direction, the zero point should be interpreted as indicating preference for Black. If the zero point is displaced from a valid value in this fashion, it would mean that the IAT identifies more persons than it should as possessing a preference for racial White.

who exhibit no behavioral preference for Whites versus Blacks." In their expectation, an average value of zero for the intercept in this regression should indicate lack of racial preference, meaning that "behavioral neutrality map[s] onto IAT neutrality" (p. 1471).

The "logic model" underlying Blanton et al.'s (2015) regression-intercept method (p.1471) can be unpacked by (a) starting from the formula for the intercept of a bivariate regression and expressing both the IAT measure and its presumed-valid zero-value predictor (X) in standard deviation (SD) units, then (b) using this logic in both the direction tested by Blanton et al. (Equation 3) and in the reverse direction (Equation 4):

$$\text{Intercept}_{\text{IAT}} = M_{\text{IAT}} - r_{\text{X-IAT}} \times M_{\text{X}} \quad (3)$$

where M_{IAT}, M_{X}, and $r_{\text{X-IAT}}$ are (respectively) mean of IAT in SD units, mean of predictor X in SD units, and the product moment correlation between X and IAT (see, e.g., Cohen, Cohen, West, & Aiken, 2003, p. 33, combining their Equations 2.4.3 and 2.4.4).

$$\text{Intercept}_{\text{X}} = M_{\text{X}} - r_{\text{X-IAT}} \times M_{\text{IAT}} \quad (4)$$

Equations 3 and 4 can be solved to find values of M_{X} and M_{IAT} that will produce zero intercepts in both directions of regression by (a) setting both intercepts to 0 and (b) setting $r_{\text{X-IAT}}$ to values observed in the various data sets analyzed by Blanton et al. (2015). The solutions will yield values for M_{IAT} and M_{X} that should produce the desired zero values of intercepts in both directions of regression, testing Blanton et al.'s logic model. Values of $r_{\text{X-IAT}}$ for the 37 data sets in Blanton et al.'s Table 6 ranged from $r = .07$ to $r = .53$. Using either of those extreme values or any values between those, the simultaneous-equation solution is that both M_{X} and M_{IAT} must equal zero. That is, values of zero for both M_{X} and M_{IAT} allow zero intercepts to be observed in both directions. Only when $r_{\text{X-IAT}}$ approaches 1.0 can zero intercepts in both directions be

observed with nonzero values of M_X and M_{IAT}, but in this case, nonzero values of the two means must be numerically equal.

Data (generously provided by Hart Blanton) for the 37 regression analyses summarized in Blanton et al.'s (2015) Table 6 were used to compute individual-study intercepts for both directions of regression. In the direction reported by Blanton et al. (regression of IAT on predictor), the weighted average intercept in SD units was 0.51, not at all close to zero. Applying Blanton et al.'s logic, 0.51 is the mean IAT score (corresponding approximately to an IAT *D* measure of 0.20) that one expects to observe among individuals who have no explicit attitude preference for Whites relative to Blacks.

Applying the regression method in the reverse direction produced a weighted average intercept of −0.01, which calls for interpretation (applying the same logic) as the mean explicit race attitude that one expects to observe among individuals who exhibit no IAT preference for Whites versus Blacks. Applying Blanton et al.'s logical model, this very close-to-zero result indicates that the IAT's zero point *is* located at an appropriate rational-zero value.

This juxtaposition of two mutually inconsistent conclusions from regression analyses computed in both directions from the same data set is, in actuality, not paradoxical. The statistics of regression intercepts oblige that, unless a regression involves two perfect measures (i.e., both test–retest reliabilities = 1.0) and a perfect correlation ($r_{X\text{-}IAT}$ = 1.0) between the two measures, the two intercepts will not be identical when the direction of regression is reversed. The data chosen by Blanton et al. were very far from meeting either the reliability criterion of perfection or the correlation criterion of perfection, obliging the conclusion that the reasoning described as

the logic explaining their choice of method was not consistent with the mathematics of bivariate regressions conducted with imperfect measures.[4]

Evidence for Construct Validity of IAT-Measured Implicit Self-Esteem

IAT measures of self-esteem do not correlate highly either with self-report measures of self-esteem or with other implicit measures of self-esteem (Bosson, Swann, & Pennebaker, 2000; Buhrmester, Blanton, & Swann, 2011). Explicit self-esteem measures have also been faulted for weakness of evidence for their construct validity (Baumeister, Campbell, Krueger, & Vohs, 2003; see also Krueger, Vohs, & Baumeister, 2008; Swann, Chang-Schneider, & McClarty, 2007). The present data afforded an opportunity to examine validity of both IAT and self-report measures of self-esteem in studies of a theory (BIT) that predicts correlations involving self-esteem measures. The meta-analysis's data provided stronger evidence, in the form of larger observed effect sizes for Tests 1 and 2 (of both the 4-test method and the within-study meta-analysis method), of validity for IAT-measured implicit self-esteem than for self-report measures of explicit self-esteem. In addition, the finding that balance–congruity effects hold in studies involving self-esteem IATs (when these are analyzed as a separate group) provided the first such demonstration in analyses of balanced identity studies. It therefore provides some of the best evidence available for nomological validity of IAT self-esteem measures.

Even while supporting construct validity of IAT-measured implicit self-esteem, the meta-analytic results showed that this evidence with self-esteem measures was, in some instances, somewhat weaker than with IAT-measured self-associations involving attributes other than valence. A possible explanation (although one not testable in the meta-analysis) follows from the

[4] The average difference between means of the IAT measure and predictor in the 37 analyses of Blanton et al.'s Table 6 was 0.48 SD units, with IAT measures indicating greater White preference than did their predictors. If both IAT measures and their predictors are assumed to have valid zero points, this substantial difference between their means is a strong indication that the two measures do not measure identical constructs.

theorized centrality of self and valence in balanced identity theory. Valence is so extensively connected to identity-relevant concepts in BIT's Social Knowledge Structure (see Figure 1 in Main Text) that its associations with group and attribute concepts should have more added associative influences than do non-valence attributes. Balanced identity studies involving novel concepts that have had no chance to develop associations other than those that are experimentally established may provide an opportunity to obtain stronger confirmation of predictions involving valence associations.

References

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, *4*, 1–44. doi:10.1111/1529-1006.01431

Blanton, H., & Jaccard, J. (2006). Tests of multiplicative models in psychology: A case study using the unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *113*, 155–166. doi:10.1037/0033-295X.113.1.155

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*, 1468–1481. doi:10.1037/a0038379

Bosson, J. K., Swann, W. B., Jr., & 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. doi:10.1037/0022-3514.79.4.631

Buhrmester, M. D., Blanton, H., & Swann, W. B., Jr. (2011). Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*, *100*, 365–385. doi:10.1037/a0021341

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist*, *61*, 56–61. doi:10.1037/0003-066X.61.1.56

Greenwald, A. G., Rudman, L. A., Nosek, B. A., & Zayas, V. (2006). Why so little faith? A reply to Blanton and Jaccard's (2006) skeptical view of testing pure multiplicative theories. *Psychological Review, 113*, 170–180. doi:10.1037/0033-295X.113.1.170

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Krueger, J. I., Vohs, K. D., & Baumeister, R. F. (2008). Is the allure of self-esteem a mirage after all? *American Psychologist, 63*, 64–65. doi:10.1037/0003-066X.63.1.64

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

Swann, W. B., Jr., Chang-Schneider, C., & McClarty, K. L. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist, 62*, 84–94. doi:10.1037/0003-066X.62.2.84