Article in press, *Social Psychological and Personality Science*

The Relation Between Evaluation and Racial Categorization of Emotional Faces

Jordan R. Axt

Center for Advanced Hindsight, Duke University, Durham, NC

Yoav Bar-Anan

Ben Gurion University of the Negev, Beer Sheva, Israel

School of Psychological Science, Tel-Aviv University, Israel

Michelangelo Vianello

University of Padova, Padova, Italy

Author Contact Information
Jordan Axt
Duke University
334 Blackwell St #320,
Durham, NC 27701
jordan.axt@gmail.com

Abstract

Prior research has found that indirectly measured preference for White people over Black people is positively related to categorizing angry racially ambiguous faces as Black. This past work found no evidence that directly measured racial preferences predict this racial categorization bias (RCB), suggesting that the RCB could be a unique and easily administered tool for investigating automatic evaluation and validating automatic evaluation measures. In two studies (Total *N* > 7,000), using structural equation models that account for error variance, multiple indirect evaluation measures were uniquely related to the RCB, thus bolstering their predictive validity. However, the RCB also correlated with self-reported evaluation, leaving psychologists without a robust, replicable outcome uniquely related to automatic evaluation. The lack of such an outcome hinders theoretical and practical progress in research on implicit social cognition.

Keywords: Automatic evaluation, implicit cognition, face perception, race, validity

The Relation Between Evaluation and Racial Categorization of Emotional Faces

Face perception is a key determinant of outcomes like interpersonal interactions (Marsh, Ambady, & Kleck, 2005) or hiring decisions (Riggio & Throckmorton, 1988). Intergroup biases in face perception are well-documented, particularly for race (e.g., Hugenberg, Young, Bernstein, & Sacco, 2010). One bias in face perception is the effect of emotional expression on racial categorization (Dunham, 2011; Hugenberg & Bodenhausen, 2004; Hehman, Ingbretsen, & Freeman, 2014; Hutchings f& Haddock, 2008): people with more negative evaluations of Black people show a racial categorization bias (RCB) where they are more likely to categorize racially ambiguous angry faces as Black than White.

An interesting aspect of the relation between evaluation and the RCB is that it has only been found with indirectly measured evaluation, like the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). Previous work, perhaps because of small sample size (e.g., 33% power at detecting a correlation of $r = .20$), found no evidence that self-reported evaluation is related to the RCB (Hugenberg & Bodenhausen, 2004), and did not report comparisons between correlations of the RCB with direct versus indirect measures of racial evaluation. The distinction between automatic and deliberate evaluation is a central tenet of dual-process attitude models (Gawronski & Bodenhausen, 2011; Greenwald & Banaji, 1995), which assume two constructs. The implicit construct reflects automatic processes that can influence behavior unintentionally and perhaps without awareness. The explicit construct reflects deliberate processes. Vast research has adopted this perspective to examine concepts like interpersonal behavior (Dovidio, Kawakami, & Gaertner, 2002; Turner, Hewstone & Voci, 2007) or self-esteem (Ziegler-Hill, 2006).

Common measures of automatic evaluation are indirect: evaluation is inferred from behavior not requiring deliberate evaluation (for review, Gawronski & De Houwer, 2014). Self-report questionnaires are typically used to measure deliberate evaluation. Progress on dual-process models depends, in part, on the validity and reliability of the measures used to assess their constructs. Without valid measures, research that highlights the usefulness of dual-process theories and improves understanding of human cognition is limited. There is strong evidence that direct and indirect measures of evaluation are related to evaluation; direct and indirect attitude measures correlate (Bar-Anan & Nosek, 2014), vary by attitude object (Nosek, 2005), and load on distinct constructs (Bar-Anan & Vianello, 2018).

The best evidence that indirect measures capture a construct not tapped by direct measures are studies that found superiority for indirect measures in predicting outcomes that are plausibly more sensitive to automatic evaluation. Common examples come from uncontrolled behavior during outgroup interactions (e.g., Dovidio, Kawakami & Gaertner, 2002; Dovidio et al., 1997). However, these studies relied on a range of outcomes, such as speech hesitation (McConnell & Leibold, 2001) or ratings of interaction quality (Gonsalkorale, von Hippel, Sherman, & Klauer, 2009), and produced conflicting results. For instance, indirect measures of evaluation predicted seating distance in Bessennoff and Sherman (2000) but not in McConnell and Leibold (2001). Moreover, other work has found differences in the contexts where such effects arise; for example, Hofmann et al. (2008) found that indirect measures of evaluation only predicted speech-related behavior under cognitive load. Finally, measuring such behavior is resource-intensive. It would be difficult to use uncontrolled behavior to learn more about discrepancies between automatic and deliberate evaluation or for improving automatic evaluation measures. The RCB presents such an opportunity.

If the RCB is related only to supposed measures of automatic evaluation, it may clarify when only automatic evaluation predicts behavior. Knowing that the RCB is related only to automatic evaluation could spur investigations into what common factors shape the RCB and automatic but not deliberate evaluation. In addition, if the RCB is only related to automatic evaluation, then a stronger association between the RCB and certain indirect measures suggests that such measures do a superior job assessing the implicit construct. Researchers could use the RCB as an outcome to compare the validity of different indirect measures or different treatments of the same measure (e.g., comparing scoring procedures). Such work does not depend on the RCB being a "process-pure" measure of automatic evaluation, or that the RCB assesses more implicit processes than other measures, like the IAT. Rather, the goal is to identify an outcome reliably related to the construct of automatic evaluation and reliably *not* related to deliberate evaluation. Identifying an outcome consistently related to automatic but not deliberate evaluation would accelerate theoretical and practical progress in implicit social cognition.

The Present Research

Although several studies found a relation between the RCB and indirectly measured evaluation (Dunham, 2011; Hehman, Ingbretsen, & Freeman, 2014; Hugenberg & Bodenhausen, 2004; Hutchings & Haddock, 2008), only Hugenberg and Bodenhausen (2004) examined the relation of the RCB with directly measured evaluation (finding no relation). However, this work had several limitations, such as small samples (*N*'s < 60) and results that were not particularly robust (e.g., the statistical significance of the relationship between the IAT and the RCB was $p = .049$ in Study 1). Moreover, inferences were based on least squares linear regression, which fails to correct for measurement reliability (Westfall & Yarkoni, 2016). Finally, previous studies used the IAT as the indirect measure, excluding other supposed measures of automatic evaluation.

This work used large samples, improved analyses, and multiple indirect measures to then test the relation between automatic and deliberate evaluation with the RCB.[1]

Study 1

Method

Participants

Participants in both studies were volunteers from the Project Implicit research pool who reported race other than Black. The study's pre-registration (https://osf.io/nast8/) outlines a sequential analysis approach (Sagarin, Ambler, & Lee, 2014), which required using $p < .0208$ as a threshold to keep the type I error rate at 5% for the test determining whether to continue data collection (the *p*-value in both study pre-registrations was computed slightly inaccurately). Pre-registered analyses were not sufficiently detailed, and we were unsure about the exact criteria for stopping data collection, deciding to run the study until the maximum planned sample size. As a result, experimental designs and data treatment are pre-registered but not analysis plans.

Study 1 had a target sample size of 4,500, which provided greater than 97% power for detecting a correlation of $r = .10$. In total, 4,691 participants ($M_{\text{Age}} = 33.3$, $SD = 14.7$, 62.4% female, 71.2% White) completed at least the racial categorization task. After exclusions, 3,124 participants were included in SEM and linear regression analyses. We report all data exclusions and measures. All materials and data can be accessed at https://osf.io/7tcwa/.

Measures

Racial Categorization. The task used the same computer-generated images as Hugenberg and Bodenhausen (2004).[2] In a task described as a "perceptual judgment task," the

[1] The online supplement (https://osf.io/7jvn6/) reports earlier studies with variations of the racial categorization task and evaluation measures. These studies were less reliable due to modest samples and complex designs, but results were similar to those reported here.

instruction was: “For each face you see, try to judge whether the face looks more like a White or a Black person. Please make these judgments as quickly as possible.” See online supplement for full instructions text.

Participants categorized each face in a random order. Each of the 15 faces was shown once with a happy and once with an angry expression. Participants were excluded from analysis if they provided the same response on all trials, and a large percentage (20%) of participants did.

Indirect Evaluation Measures. Participants were randomly assigned to complete one of three indirect racial evaluation measures: an IAT (Greenwald et al., 1998), an Affect Misattribution Procedure (AMP; Payne, Cheung, Govorun & Stewart, 2005), or an evaluative priming task (EPT; Fazio et al., 1995). Each measure used the same image stimuli, and the IAT and EPT used the same positive and negative words (Gawronski, Gast, & De Houwer, 2015).

The IAT followed the design recommended by Nosek, Greenwald, and Banaji (2005), and measured the strength of association between categories "White people" and "Black people" and the concepts "Negative" and "Positive." Responses were scored by the *D*600 algorithm (Greenwald, Nosek, & Banaji, 2003), with higher scores reflecting more positive associations with White versus Black people. Participants were excluded from analyses if more than 10% of critical block responses were faster than 300ms (Greenwald et al., 2003; 1.3% of IAT scores).

On each AMP trial, participants observed a White or Black face for 100ms, followed by a blank screen for 100ms, followed by a Chinese symbol for 100ms, followed by a display of gray "noise", presented until participants rated the symbol as more pleasant or unpleasant than average. Following the same procedure as Moran, Bar-Anan, and Nosek (2017), participants completed three blocks of 40 trials (20 trials for each prime category). The final AMP score was

[2] The original article using Study 1 stimuli (Hugenberg & Bodenhausen, 2004) reported pretesting faces to be “difficult” to categorize by race when emotionally neutral.

the percentage of 'Pleasant' responses following White versus Black primes. Participants were excluded from analyses if 95% or more of critical trials had the same response (Bar-Anan & Nosek, 2014; 16.0% of AMP scores).

On each EPT trial, a White or Black face appeared for 200ms, followed by a target word presented until participants categorized it as “Negative” or “Positive.” If participants responded incorrectly, a red X appeared. The EPT consisted of three, 60-trial blocks, following the same procedure as Moran et al. (2017). EPT performance was scored using the same method as Bar-Anan and Nosek (2014), with higher scores indicating greater facilitation of White faces for identifying positive words and Black faces for identifying negative words (see online supplement for full scoring procedure). Participants were excluded if more than 40% of trials were errors (Bar-Anan & Nosek, 2014; 2.9% of EPT scores).

Direct Evaluation. Participants completed five items assessing self-reported preference between Black and White people. Participants reported how much they preferred Black vs. White people (-3 = Strongly prefer Black people, +3 = Strongly prefer White people), then how much they liked White and Black people separately (1= Strongly Dislike, 7= Strongly like), and finally used a slider to indicate how positive or negative they felt towards White and Black people separately (0=Extremely negative, 100= Extremely positive). A composite measure was made by creating difference scores from the liking and thermometer slider items (more positive values indicated greater preference for White people), then standardizing and averaging the relative explicit preference item and the two difference scores (ɑ = .82). We only included participants who responded to all self-report items.

Procedure

Participants completed the race categorization task followed by randomly ordered direct and indirect evaluation measures.

Results

Descriptive Statistics

Average scores on each evaluation measure are presented in Table 1. In the racial categorization task, participants had higher rates of categorizing happy (60.5%) versus angry (59.05%) faces as Black, $t(3749) = 5.54$, $p < .001$ $d = .09$. These analyses were not reported in Hugenberg and Bodenhausen (2004). Other related work either found that angry versus happy faces were more likely to be categorized as Black (Dunham, 2011; Hutchings & Haddock, 2008), or found no emotional differences in racial categorization (Hehman, Ingbretsen & Freeman, 2014). However, the large samples sizes used here were more sensitive to small effects that would likely have been reported as nulls in Hugenberg and Bodenhausen (2004) or Hehman, Ingbretsen, and Freeman (2014), where *N*'s < 58.

Correlations of Racial Categorization with Direct and Indirect Measures

As in Hugenberg and Bodenhausen (2004), we computed the RCB as the difference between the percentage of angry versus happy faces judged as Black, with higher values indicating greater likelihood of perceiving angry versus happy faces as Black. Table 2 shows positive correlations between all evaluation measures and the RCB (*r*'s > .119, *p*'s <.001). Self-reported racial evaluation was positively and reliably related to the RCB (*r*'s > .17, p's < .001). Effect sizes suggested that similar associations would have likely gone undetected in prior work, as Study 1 of Hugenberg and Bodenhausen (2004) had only 17% power to detect the largest correlation between the RCB and self-reported racial evaluation found in Study 1.

Table 1. Descriptive statistics of Study 1 and Study 2 measures.

Study 1	
Measure (*N*)	*M* (SD)
RCB (*N* =3750)	-0.01 (0.16)
Explicit Preference (*N* =4418)	0.27 (0.89)
Thermometer Difference Score (*N* =4418)	-0.18 (27.61)
Liking Difference Score (*N* =4418)	-0.01 (1.02)
IAT (*N* =1478)	0.31 (0.38)
AMP (*N* =1219)	-0.03 (0.20)
EP (*N* =1353)	0.11 (0.42)
Study 2	
Measure (*N*)	*M* (SD)
RCB (*N* =3793)	-0.13 (0.22)
Explicit Preference (*N* =3750)	0.32 (0.83)
Thermometer Difference Score (*N* =3750)	1.02 (26.64)
Liking Difference Score (*N* =3750)	0.03 (0.98)
Face Pleasantness Difference Score (*N* =3753)	-0.12 (0.65)
ST-IAT (*N* =1835)	-0.06 (0.28)
AMP (*N* =1604)	-0.03 (0.19)

Note. All measures scored such that higher values mean more negative evaluations of Black (versus White) people. RCB = Racial categorization bias; IAT= Implicit Association Test; AMP = Affect Misattribution Procedure; EP = Evaluative priming task; ST-IAT = Single-Target Implicit Association Test.

Table 2. Correlations between measures in Studies 1-2

Study 1

IAT Condition

Measure	RCB	Direct Evaluation
Direct Evaluation	.23	
IAT	.17	.21

AMP Condition

Measure	RCB	Direct Evaluation
Direct Evaluation	.18	
AMP	.26	.36

EP Condition

Measure	RCB	Direct Evaluation
Direct Evaluation	.17	
EP	.12	.18

Study 2

ST-IAT Condition

Measure	RCB	Direct Evaluation	Pleasant Rating
Direct Evaluation	.14		
Pleasant Rating	.19	.45	
ST-IAT	.10	.14	.19

AMP Condition

Measure	RCB	Direct Evaluation	Pleasant Rating
Direct Evaluation	.17		
Pleasant Rating	.21	.43	
AMP	.21	.28	.45

Note: Direct evaluation = Self-report racial attitude composite variable; Pleasant rating = Difference score from pleasantness ratings for White versus Black faces. All correlations reliable at $p < .01$.

Replication of Original Analyses

For each indirect measure, we repeated Hugenberg and Bodenhausen's (2004) IAT analyses. The first used least squares linear regression to predict the RCB from mean-centered values of the direct and indirect measures, as well as their interaction. Replicating Hugenberg and Bodenhausen (2004), each indirect measure reliably predicted the RCB (β's > .09, *t*'s > 2.88, *p*'s < .004; see Table 3). Unlike Hugenberg and Bodenhausen's results, self-reported racial attitudes also reliably predicted the RCB (β's > .09, *t*'s > 2.98, *p*'s < .003). None of the interactions between direct and indirect measures were reliable (using the adjusted *p*-value criterion).

SEM Analyses

To compare the predictive strength of the self-reported evaluation and each indirect measure, we used structural equation modeling (SEM) to control for measurement reliability, as least squares linear regression analyses may inflate Type I error rates in claims of incremental predictive validity (Westfall & Yarkoni, 2016).

We specified and estimated three models, one for each indirect measure, predicting the RCB from the indirect and direct measures. We used three indicators for the implicit construct (measured with the indirect measure) and three for the explicit construct (relative preference item, liking difference score, thermometer difference score). The AMP (*a* = .79) and EPT's (*a* = .49) three indicators were the scores computed from each task's three blocks. For the IAT, the 60 critical trials in blocks 3-4 were matched with the 60 critical trials blocks 6-7, then divided into three parcels of 20 trials (first 20 trials into the first parcel, etc.) with separate *D* scores calculated for each parcel (*a* = .65). We tested whether the 2-factor solution better fit the data

Table 3. Results of linear regressions predicting the RCB from indirect and direct measures in Studies 1-2.

Study 1					
IAT Condition (*N* = 1132; R^2 = .067)					
Term	B	*SE*	*t*	*β*	*p*
IAT	.06	.01	4.52	.13	<.001
Direct Evaluation	.04	.01	6.73	.20	<.001
IAT * Direct Evaluation	.01	.01	0.53	.02	.600
AMP Condition (*N* = 958; R^2 = .081)					
Term	B	*SE*	*t*	*β*	*p*
AMP	.20	.03	6.79	.23	<.001
Direct Evaluation	.02	.01	2.98	.10	.003
AMP * Direct Evaluation	-.04	.02	-2.17	-.07	.030
EP Condition (*N* = 1034; R^2 = .034)					
Term	B	*SE*	*t*	*β*	*p*
EP	.03	.01	2.88	.09	.004
Direct Evaluation	.03	.01	4.72	.15	<.001
EP * Direct Evaluation	-.003	.01	-0.22	-.01	.826
Study 2					
ST-IAT Condition (*N* = 1667; R^2 = .027)					
Term	B	*SE*	*t*	*β*	*p*
ST-IAT	.06	.02	3.25	.08	.001
Direct Evaluation	.03	.01	5.34	.13	<.001
ST-IAT * Direct Evaluation	-.02	.02	-1.23	-.03	.218
AMP Condition (*N* = 1442; R^2 = .060)					
Term	B	*SE*	*t*	*β*	*p*
AMP	.21	.03	6.54	.17	<.001
Direct Evaluation	.03	.01	4.81	.13	<.001
AMP * Direct Evaluation	-.04	.03	-1.65	-.04	.099

Note. SE = Standard error.

in comparison with a solution where all indicators loaded onto one factor. The two-factor solutions yielded a better fit (AMP: $\Delta\chi^2$=77.6, *p*<.001); EP: $\Delta\chi^2$=43.5, *p*<.001; IAT: $\Delta\chi^2$=192.6, *p*<.001). Next, we estimated the two-factor models with the criterion variable (see Figure 1 for the "AMP" model).

Each model showed acceptable fit (AMP: $\chi^2(12) = 24.62$, $p = .017$, *Comparative Fit Index* (*CFI*) = .92, *Root Mean Square Error of Approximation* (*RMSEA*) = .033 95% *Confidence Interval* (*CI*) = [.014, .052]; IAT: $\chi^2(12) = 18.76$, $p = .095$, *CFI* = .98, *RMSEA* = .022 95% *CI* = [.000, .041]; EP: $\chi^2(12)=9.12$, $p = .693$, *CFI* = 1.00, *RMSEA* = .000 95% *CI* = [.000, .025]). Table 4 provides regression estimates, tests of difference from zero, and bootstrapped bias-corrected confidence intervals (BCCIs) around the standardized regression estimate. Direct and indirect evaluation measures reliably predicted the RCB with one exception: the self-report variable was not a reliable predictor in the AMP model. Pairwise comparisons showed that only the AMP predicted the RCB better than the direct measure.

To test for differences in predictive strength among indirect measures, classical pairwise comparisons of standardized differences between parameter values could not be run because each indirect measure was collected on an independent sample. Hence, we ran the Wald test using the unstandardized estimate of the path between the AMP and the RCB as the effect of interest, and compared it toward the unstandardized estimate of the paths between the IAT, the EP and the RCB, using the standard error of the AMP estimate as a denominator. The predictive power of the AMP was not significantly higher than the EP or IAT (respectively, *Zep*=.36, *p*=.36, *Ziat*=.27, *p* =.39).

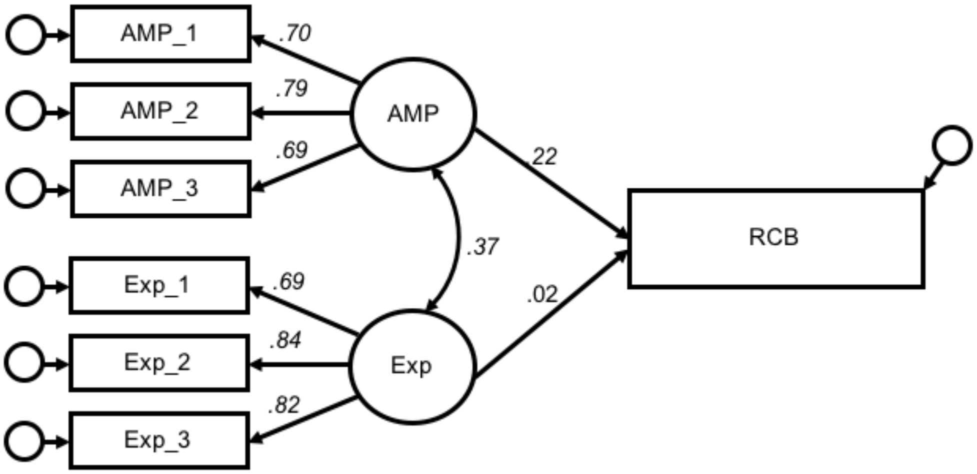

Figure 1. The AMP model for testing the relative predictive power of direct and indirect measures. Significant estimates are displayed in Italics. Only standardized estimates are shown. See text for model fit statistics.

Table 4. Study 1: Estimates of regression paths between direct and indirect predictors of the RCB.

Model	Regression weight			Estim.	S.E.	Test of the difference from zero		Std estim.	95% BC-CI*		Indirect-Direct comparison		R^2
						Z	*p*		LB	UB	*Z*	*p*	
EP	Exp	→	RCB	.033	.010	3.357	<.001	.15	.06	.24			.05
	EP	→	RCB	.044	.021	2.134	.033	.13	.01	.20	-.38	.65	
IAT	Exp	→	RCB	.049	.010	5.142	<.001	.22	.15	.30			.09
	IAT	→	RCB	.045	.012	3.857	<.001	.16	.08	.23	-.22	.39	
AMP	Exp	→	RCB	.004	.011	.484	.628	.021	-.08	.12			.05
	AMP	→	RCB	.048	.011	4.536	<.001	.223	.14	.31	2.75	.01	

*Bootstrap based on 1000 samples.

Discussion

Parallel analyses of those in Hugenberg and Bodenhausen (2004) found that indirect measures of racial evaluation predicted the RCB, though self-reported evaluation did as well Indirect measures of evaluation and self-report each uniquely predicted the RCB, with the exception of self-report when including the AMP.

We extended these findings in Study 2. We replaced the EPT and IAT with a Single-Target IAT (ST-IAT; Bluemke & Friese, 2008). In addition, a limitation of Study 1 was that only the indirect measures of evaluation included facial stimuli. To address the possibility that Study 1 results were due to the indirect measures assessing facial instead of automatic evaluation, Study 2 included a measure where participants directly rated the pleasantness of the facial stimuli.

Study 2

Method

Participants

Planned sequential analysis in the pre-registration (https://osf.io/htkpz/) reduced the critical *p*-value to .021. We targeted a sample size of 3,600 participants, providing 97% power for detecting an $r = .10$ correlation. In total, 3,981 participants ($M_{\text{Age}} = 33.4$, $SD = 14.6$, 61.0% female, 75.8% White) completed at least the racial categorization task. After exclusions, 3,109 participants entered SEM and linear regression analyses.

Measures

Racial Categorization. The task used different stimuli as Study 1 (Hehman, Ingbretsen, & Freeman, 2014) in hopes of lowering the percentage of participants who categorized all the stimuli to the same race. This, however, decreased the validity of inferences made from comparisons between the studies.

In a random order, participants saw 15 racially ambiguous happy faces (selected randomly for each participant from a pool of 29) and 15 racially ambiguous angry faces (from a pool of 89).[3] We excluded participants who gave the same response to all trials (4.7% of participants who completed the task).

Indirect Evaluation Measures. Participants were randomly assigned to complete either an AMP or ST-IAT. The AMP had the same design, scoring, and exclusion criteria (15.1%) as Study 1. The ST-IAT followed the design recommended by Bluemke and Friese (2008) and measured the strength of association between the category "Black people" and the concepts "Negative" and "Positive," using the same stimuli as the IAT in Study 1. Responses were scored by the *D*600 algorithm, with higher scores indicating more positive versus negative associations with Black people. Participants were excluded from analyses if more than 10% of responses were faster than 300ms (1.8% of ST-IAT scores).

Direct Evaluation Measures. Participants reported racial attitudes identically to Study 1.

Face Rating. Participants rated each of the 24 faces used in the AMP (and the same Black faces as the ST-IAT) on pleasantness (1 = Very unpleasant, 7 = Very pleasant). The outcome score was the difference between ratings of White vs. Black faces (higher values indicated pro-White ratings).

Procedure

Participants completed the race categorization task followed by randomly ordered direct evaluation measure, indirect evaluation measure, and face pleasantness rating.

[3] The article introducing Study 2 stimuli (Hehman, Inbgretsen & Freeman, 2014) reported no pretesting, but used facial morphs of computer-generated White and Black faces.

Results

Descriptive Statistics

Overall scores (Table 1) were similar to Study 1. Again, the racial categorization task revealed higher rates of categorizing happy (55.3%) versus angry (42.1%) faces as Black, $t(3792) = 37.11$, $p < .001$ $d = 60$. The AMP, ST-IAT and direct face rating also showed pro-Black scores on average.

Correlations Between Racial Categorization with Direct and Indirect Measures

Table 2 shows positive correlations between evaluation measures, the direct face ratings, and the RCB (*r*'s > .096, *p*'s <.001). As in Study 1, self-reported evaluation was weakly but positively related to the RCB (*r*'s > .139, *p*'s < .001).

Replication of Original Analyses

Table 3 presents the results of least squares linear regression models predicting the RCB using the same approach as Study 1. Both indirect measures reliably predicted the RCB (*β*'s > .08, *t*'s > 3.25, *p*'s < .001), as did the direct measure (*β*'s > .13, *t*'s > 4.82, *p*'s < .002). None of the interactions between direct and indirect measures were reliable.

We ran the same analyses with the addition of direct face rating (mean-centered) as a predictor. In each model, the indirect and direct measures of evaluation, as well as face pleasantness ratings, predicted the RCB (*β*'s > .06, *t*'s > 2.48, *p*'s < .013; see online supplement). None of the two-way or three interactions between measures were reliable.

SEM Analyses

The SEM models for the AMP (Figure 2) and the ST-IAT (Figure 3) were the same as Study 1, with the addition of an observed variable for the direct face rating.

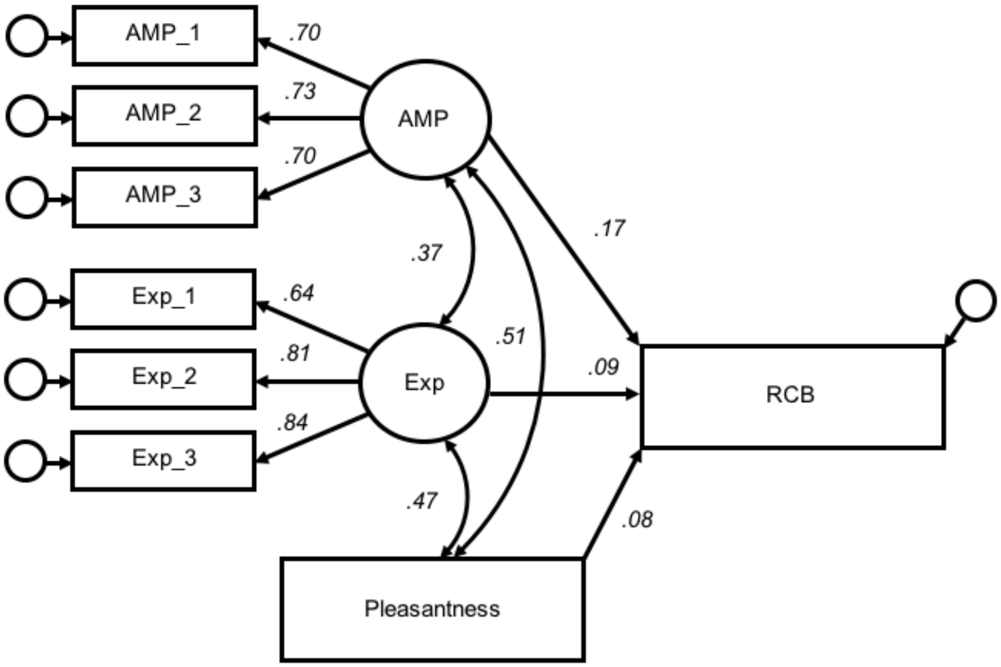

Figure 2. The AMP model for testing the relative predictive power of direct and indirect measures controlling for perceived pleasantness of the faces. Significant estimates are displayed in Italics. Only standardized estimates are shown. $\chi^2(16)=51.21$, *p* <.001, Comparative Fit Index =.99, Root Mean Square Error of Approximation =.04, 95% Confidence Interval [.03, .05].

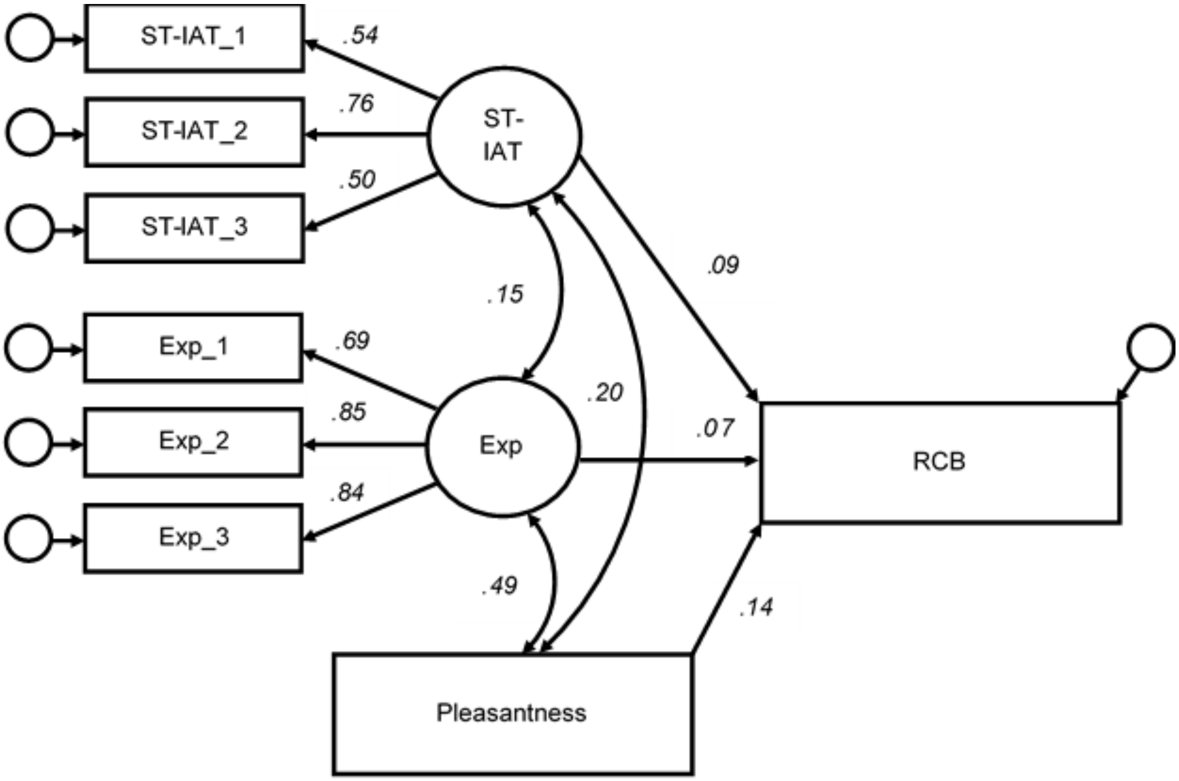

Figure 3. The ST-IAT model for testing the relative predictive power of direct and indirect measures controlling for perceived pleasantness of the faces. Significant estimates are displayed in Italics. Only standardized estimates are shown. $\chi^2(16)$=35.40, p=.004, Comparative Fit Index =.99, Root Mean Square Error of Approximation=.03, 95% Confidence Interval [.02, .04].

Table 5. Estimates of regression paths between direct and indirect predictors of the RCB.

Model	Regression weight			Estim.	S.E.	Test of the difference from zero		Std estim.	95% BC-CI*		R^2
						Z	*p*		LB	UB	
ST-IAT	Exp	→	RCB	.02	.01	2.68	.007	.07	.003	.13	.05
	ST-IAT	→	RCB	.08	.03	2.97	.003	.09	.03	.15	
	Pleas.	→	RCB	.05	.01	4.81	<.001	.14	.09	.20	
AMP	Exp	→	RCB	.03	.01	2.67	.002	.09	.02	.16	.07
	AMP	→	RCB	.24	.05	4.46	<.001	.17	.08	.25	
	Pleas.	→	RCB	.03	.01	2.33	.02	.08	-.001	.15	

Note: *Bootstrap based on 1000 samples.

The indicators of the AMP (*a* = .76) and self-report (*a* = .81) were computed as in Study 1. The indicators of the ST-IAT (*a* = .65) were computed with the same method used for Study 1's IAT. As in Study 1, the 2-factor solutions yielded a better model fit (AMP: $\Delta\chi^2$=84.6, *p*<.001); ST-IAT: $\Delta\chi^2$=572.2, *p*<.001).

Results of the full models are detailed in Table 5. All direct and indirect measures reliably predicted the RCB. As in Study 1, the AMP was a stronger predictor than direct evaluation (*Z* = 3.62, *p* < .001), whereas the ST-IAT was not (*Z* = 1.96, *p* = .06). The predictive power of the AMP was significantly higher than that of the ST-IAT (*Z* = 3.2, *p*<.001).

General Discussion

Bias in categorization of angry versus happy faces as White or Black (the RCB) was independently predicted by both direct and indirect measures of racial evaluation. Direct and indirect measures predicted the RCB when tested in isolation and when controlling for the other, using least squares linear regression and SEM. There was no consistent evidence that deliberate evaluation was inferior to automatic evaluation in predicting the RCB.

Relation with Automatic Evaluation

The present findings replicate and extend previous work linking the RCB to automatic evaluation (Dunham, 2011; Hugenberg & Bodenhausen, 2004; Hutchings & Haddock, 2008). Previous results were limited to one indirect measure of evaluation (the IAT), whereas the present work provides conceptual replications with the ST-IAT, AMP, and EPT. Across indirect measures, the present research consistently found reliable relations with RCB that were smaller than the relation found in the original works, suggesting that the relation between automatic evaluation and the RCB might be smaller than previously estimated.

Further, by using SEM, the present work showed that supposed measures of automatic evaluation uniquely predicted the RCB. This finding is notable because the SEM analysis cleans for error variance and, to our knowledge, provides the first use of such analyses to show indirect measures of evaluation independently predict criterion (c.f., Brick & Lai, 2018). That each indirect measure showed incremental predictive validity beyond direct measures of racial evaluation bolsters the claim that the RCB is uniquely related to a construct shared among these measures (i.e., automatic evaluation). While these results are consistent with the notion that each indirect measure is a valid assessment of automatic evaluation, they cannot rule out the possibility that each measure differentially assesses a unique construct related to the RCB, or that all measures capture a shared theoretical construct *other* than automatic evaluation (e.g., shared method variance). However, this latter interpretation would be inconsistent with prior work using a multi-method multi-trait design (Bar-Anan & Vianello, 2018), wherein the indirect measures used here were much more related to a shared theoretical construct of implicit racial attitudes than to a shared methodological factor.

Relation with Deliberate Evaluation

The present findings provide strong evidence that deliberate evaluation is also related to the RCB, as self-reported evaluation predicted the RCB in nearly all conditions. This association was also found in studies we report in the supplement (https://osf.io/7jvn6/) that used several versions of the racial categorization task, one of which was a lab study conducted in Israel with facial stimuli of Israeli ethnic groups. That study contributes to our confidence in the generality of our results, despite their discrepancy with results of the only other two published studies (Hugenberg & Bodenhausen, 2004) that included deliberate evaluation. We speculate that the previous studies did not have sufficient statistical power to detect that relation (all *N*s < 60).

The present data suggest that, contrary to prior work, the RCB does not reflect a psychological process related only to automatic evaluation. Identifying an outcome that is only related to automatic evaluation is highly desirable, as it would allow researchers the opportunity to further understand the distinction between automatic and deliberate evaluation. Moreover, an outcome uniquely related to automatic evaluation would be useful as a criterion in efforts to improve automatic evaluation measures.

Earlier findings seemed plausible because the RCB is likely a result of unintentional processes that use the combination of emotional facial expressions and individual's racial evaluations to make non-affective perceptual judgments based on non-evaluative visual features. However, the present results are also reasonable because intentional evaluation of racial groups could have unintentional effects. For instance, people might think about their racial attitudes while making categorizations, and those evaluations could have unintended impacts on judgment. Another possibility is that people are aware of the potential influence of their automatic evaluations on racial categorization, and try to correct for this influence deliberately. Such correction processes are likely a source of variance in deliberate but not automatic evaluation.

These studies advance our understanding of the RCB by illustrating that direct and indirect measures of racial evaluation are related to the behavior. Our finding that deliberate evaluation *is* related to the RCB does not invalidate the RCB as a topic of research; rather, these data suggest that the RCB is an outcome related to racial attitudes more generally, with unique associations with both direct and indirect measures of those attitudes. This result may have important implications for understanding how racial biases in face perception develop. Knowledge about the RCB could benefit from future work that investigates the shared or distinct factors that lead to associations with direct and indirect racial evaluation. In addition, future studies should investigate the possible causal relationship between automatic evaluation and the RCB. If the RCB is partly caused by automatic evaluation, it suggests new paths for addressing racial biases in face perception.

Missing: An Outcome Measure for Automatic Evaluation Research

The RCB's relation with deliberate evaluation leaves psychologists without an easily replicable outcome related to automatic but not deliberate evaluation. A popular outcome previously linked to automatic and not deliberate evaluation is nonverbal behavior (e.g., McConnell & Leibold, 2001). However, as reviewed earlier, such studies are highly resource intensive, and have found conflicting evidence regarding what behaviors or contexts result in an association with indirect but not direct evaluation measures.

Other studies identified either sub-populations whose behavior is uniquely related to indirect measures or contexts where behavior is related only to indirect measures. For example, undecided voters' intended political behavior was unrelated to self-reported attitudes but was predicted by an indirect measure of evaluation (an SC-IAT; Galdi, Arcuri, & Gawronski, 2008). However, subsequent investigations of undecided voters found that direct and indirect measures of evaluation were both related to behavioral intentions and actual behavior (Friese et al., 2012). In other work, participants' dieting choices under cognitive load were uniquely predicted by an indirect measure of evaluation (an IAT; Friese, Hofmann, & Wanke, 2008), but a previous study using another cognitive load manipulation found that an IAT failed to predict health-related behavior (Scarabis, Florack, & Gosejohann, 2006).

Still other work has identified outcomes that were related to indirect but not direct measures of evaluation (e.g., Gawronski, Geschke, & Banse, 2003; Holland, de Vries, Hermsen,

& Van Knippenberg, 2012; Lane, Goh, & Driver-Linn, 2012), but to our knowledge, none of these have been replicated directly or conceptually using other indirect measures. Despite over 20 years of research, the field has yet to identify an outcome that is consistently and uniquely predicted by indirect evaluation measures and not by direct evaluation measures. Such an outcome measure would simplify and facilitate the empirical investigation of automatic evaluation and its measures.

The Superiority of the AMP

We found suggestive evidence that the AMP was more strongly related to the RCB than other indirect measures. Further, unlike other indirect measures, the AMP was a stronger predictor of the RCB than direct racial evaluation (similar results were found in Study S4, reported in the online supplement). These findings might suggest that the AMP is a superior measure of automatic evaluation, which would align with previous evidence that the AMP has better predictive validity than the IAT or EPT (Payne, Govorun, & Arbuckle, 2008). However, it is notable that the relation between self-report and the RCB lessened when the AMP was added as a predictor, suggesting that some of the AMP's predictive validity comes from sensitivity to deliberate processes. This possibility is compatible with evidence that the AMP, unlike other indirect measures, is equally related to direct and indirect evaluation measures (Bar-Anan & Vianello, 2018). Because the RCB is related to both deliberate and automatic evaluation, it is difficult to know whether good predictive validity is evidence of superior measurement for automatic processes or for deliberate processes missed by the self-report measure.

The difficulty in interpreting the AMP's superiority is emblematic of our conclusions regarding the RCB. Our results are compatible with the conclusion that the RCB is related to automatic evaluation, but are not definitive because the RCB is also related to deliberate evaluation. Even when an indirect measure outperforms self-reported evaluation in predicting the RCB, there is still the threat that the indirect measure is superior *because* of its heightened sensitivity to deliberate rather than automatic processes.

In summary, the present results are compatible with the possibility that automatic evaluation is an important construct that is related to the RCB, and is captured by the IAT, AMP, EPT, and ST-IAT. The results are also compatible with the possibility that the AMP is superior to other indirect measures in capturing automatic evaluation. However, for more definitive conclusions, researchers are still missing an easily replicable outcome that is related to automatic but not deliberate evaluation. This omission presents a conceptual and practical challenge for advancing knowledge about implicit social cognition.

References

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668-688.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264-1272.

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*(4), 329-353.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*(6), 977-997.

Brick, C., & Lai, C.K. (2018). Putting the "I" in environmentalist: Explicit (but not implicit) identity predicts pro-environmental action. *Journal of Environmental Psychology, 58*, 8-17.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62-68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, *33*(5), 510-540.

Dunham, Y. (2011). An angry = outgroup effect. *Journal of Experimental Social Psychology*, *47*(3), 668-671.

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*(6), 1013-1027.

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behavior. *British Journal of Social Psychology*, *47*(3), 397-419.

Friese, M., Smith, C. T., Plischke, T., Bluemke, M., & Nosek, B. A. (2012). Do implicit attitudes predict actual voting behavior particularly for undecided voters?. *PloS One, 7*(8), e44130.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321*(5892), 1100-1102.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model:

Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59-127.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York: Cambridge University Press.

Gawronski, B., Gast, A., & De Houwer, J. (2015). Is evaluative conditioning really resistant to extinction? Evidence for changes in evaluative judgements without changes in evaluative representations. *Cognition and Emotion*, *29*(5), 816-830.

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*(5), 573-589.

Gonsalkorale, K., von Hippel, W., Sherman, J. W., & Klauer, K. C. (2009). Bias and regulation of bias in intergroup interactions: Implicit attitudes toward Muslims and interaction quality. *Journal of Experimental Social Psychology*, *45*(1), 161-166.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 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*(2), 197-216.

Hehman, E., Ingbretsen, Z. A., & Freeman, J. B. (2014). The neural basis of stereotypic impact on multiple social categorization. *Neuroimage*, *101*, 704-711.

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations*, *11*(1), 69-87.

Holland, R. W., Vries, M. D., Hermsen, B., & Knippenberg, A. V. (2012). Mood and the attitude–behavior link: The happy act on impulse, the sad think twice. *Social Psychological and Personality Science*, *3*(3), 356-364.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342-345.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorizationindividuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168-1187.

Hutchings, P. B., & Haddock, G. (2008). Look Black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *Journal of Experimental Social Psychology*, *44*(5), 1418-1420.

Lane, K. A., Goh, J. X., & Driver-Linn, E. (2012). Implicit science stereotypes mediate the relationship between gender and academic participation. *Sex Roles*, *66*(3-4), 220-234.

Marsh, A. A., Ambady, N., & Kleck, R. E. (2005). The effects of fear and anger facial expressions on approach-and avoidance-related behaviors. *Emotion*, *5*(1), 119-124.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*(5), 435-442.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology*, *47*(6), 708-723.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565-584.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*(2), 166-180.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277-293.

Payne, B. K., Govorun, O., & Arbuckle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking?. *Cognition & Emotion*, *22*(2), 238-271.

Riggio, R. E., & Throckmorton, B. (1988). The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews. *Journal of Applied Social Psychology*, *18*(4), 331-348.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293-304.

Scarabis, M., Florack, A., & Gosejohann, S. (2006). When consumers follow their feelings: The impact of affective or cognitive focus on the basis of consumers' choice. *Psychology & Marketing*, *23*(12), 1015-1034.

Turner, R. N., Hewstone, M., & Voci, A. (2007). Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of Personality and Social Psychology*, *93*(3), 369-388.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719.

Ziegler-Hill, V. (2006). Discrepancies between implicit and explicit self-esteem: Implications for narcissism and self-esteem instability. *Journal of Personality*, *74*(1), 119-144.