Journal of Personality and Social Psychology
2015, Vol. 108, No. 4, 553–561

0022-3514/15/$12.00 http://dx.doi.org/10.1037/pspa0000016

COMMENT

Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects

Anthony G. Greenwald
University of Washington

Mahzarin R. Banaji
Harvard University

Brian A. Nosek
University of Virginia and Center for Open Science, Charlottesville, Virginia

Greenwald, Poehlman, Uhlmann, and Banaji (2009; GPUB hereafter) reported an average predictive validity correlation of $\bar{r} = .236$ for Implicit Association Test (IAT) measures involving Black–White racial attitudes and stereotypes. Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013; OMBJT) reported a lower aggregate figure for correlations involving IAT measures ($\bar{r} = .148$). The difference between the estimates of the 2 reviews was due mostly to their use of different policies for including effect sizes. GPUB limited their study to findings that assessed theoretically expected attitude–behavior and stereotype–judgment correlations along with others that the authors expected to show positive correlations. OMBJT included a substantial minority of correlations for which there was no theoretical expectation of a predictive relationship. Regardless of inclusion policy, both meta-analyses estimated aggregate correlational effect sizes that were large enough to explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons.

Keywords: Implicit Association Test, predictive validity, meta-analysis, effect size, race discrimination

Within a few years after its first publication, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) had established validity as a general method for measuring relative association strengths (Greenwald & Nosek, 2001; Greenwald, Nosek, & Banaji, 2003; see also Nosek, Greenwald, & Banaji, 2007). Less than a decade later, research using IAT measures had accumulated sufficiently to permit a meta-analysis of predictive validity that encompassed nine criterion-measure domains (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; GPUB hereafter). With some variability across the nine domains, GPUB found moderate predictive validity of IAT attitude measures, echoing preceding years' demonstrations of moderate predictive validity for self-report attitude measures (e.g., Ajzen & Fishbein, 1977; Fazio, 1990; Kraus, 1995). A new finding reported by GPUB was that racially discriminatory behavior was significantly better predicted by IAT measures ($\bar{r} = .236$) than by self-report measures ($\bar{r} = .117$).

Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013; OMBJT hereafter) reported a meta-analysis of IAT studies, limiting attention to ones using criterion measures involving behavior toward stigmatized groups, encompassing two of the nine domains covered by GPUB. Their analysis on only the Black–White race subset of studies estimated a lower mean weighted correlational effect size for the IAT ($\bar{r} = .148$ [OMBJT Table 1, data for "Black vs. White groups: Overall"]) than did GPUB ($\bar{r} = .236$ for Black–White race IATs [GPUB, Table 3]).[1] OMBJT interpreted their finding as indicating that "the IAT provides little insight into who will discriminate against whom" (p. 188).

This article's reexamination of the two meta-analyses (a) assesses the extent to which their different conclusions were due to differences in methods, (b) describes conclusions that are supported by both meta-analyses, and (c) explains how effects of the small-to-moderate magnitudes observed in the two meta-analyses are expected to have societally consequential impacts.

This article was published Online First November 17, 2014.

Anthony G. Greenwald, Department of Psychology, University of Washington; Mahzarin R. Banaji, Department of Psychology, Harvard University; Brian A. Nosek, Department of Psychology, University of Virginia and Center for Open Science, Charlottesville, Virginia.

Anthony G. Greenwald and Brian A. Nosek are officers of Project Implicit, Inc., a nonprofit corporation that has as part of its scientific mission "[T]o develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors."

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle WA 98195-1525. E-mail: agg@u.washington.edu

[1] In their article, OMBJT reported effect size values computed to two decimal places. We computed these more precisely by using the Hedges, Tipton, & Johnson, (2010) analysis method that OMBJT relied on, and we are grateful to Fred Oswald for providing access to details needed to reproduce their statistical analyses.

Method Differences Account for Differing Results

Different Criteria for Including Studies and Effect Sizes in the Two Meta-Analyses

The first published study of the IAT's predictive validity appeared in September, 2000 (Phelps et al., 2000). Starting with that study and including studies available through February 2007, GPUB identified 122 relevant reports (23 unpublished) that evaluated the predictive validity of IAT measures. GPUB categorized these into nine domains: Black versus White race, other stigmatized groups, gender/sexual orientation, consumer preferences, political preferences, personality traits, alcohol and drug use, clinical phenomena, and close relationships.

OMBJT focused on the first two of GPUB's nine domains and included 27 of the 32 reports that GPUB had categorized as concerning Black–White race or other stigmatized groups. OMBJT also included 18 reports that were published between March 2007 and December 2011, along with 1 pre-2007 report that had been missed by GPUB (Shelton, Richeson, Salvatore, & Trawalter, 2005).

The most consequential difference in method between GPUB and OMBJT was the difference in their policies for including effect sizes (IAT–criterion correlations; ICCs).[2] GPUB's inclusions were guided by the authors' stated expectations for findings, which were based mostly on two familiar hypotheses: (a) measures of attitude toward a group should predict behavior favorable or unfavorable to the group and (b) measures of a stereotype of the group should predict stereotype-consistent judgments or behavior toward members of that group. As one example, the use of an IAT measure of a stereotype, one linking Black versus White race to athletics versus academics, to predict judgments about a Black person's academic involvement was included (cf., Amodio & Devine, 2006) by GPUB, whereas the use of an IAT attitude measure of Black versus White racial preference was not expected to correlate with that same criterion measure of a Black person's relative skills at athletics versus academics, and was therefore not included. In addition to these familiar theoretical bases for including attitude and stereotype studies, additional ICCs were included by GPUB when authors offered rationale for expecting the correlation to provide evidence of predictive validity. These rationales were diverse.

OMBJT stated a broader inclusion policy: "We included any study for which an [ICC] could be computed where the criterion arguably measured some form of discrimination" (p. 177). This policy did not require a theoretical basis for expecting a positive correlation, such as attitude–behavior consistency or stereotype–judgment consistency. For example, OMBJT included two correlations from the Amodio and Devine (2006) study that GPUB had excluded, as explained just above.

In implementing their policy, OMBJT included 33 ICCs that correlated a Black–White race IAT with a criterion measure involving behavior toward or judgments about a White person.[3] As is described in this article's concluding *Discussion*, 20 of those 33 ICCs were not reported by authors in the original articles or reports. The remaining 13, although included in reports, were described as not expected to show predictive validity correlations. There is no intention to suggest here that OMBJT's strategy was unwise. This article later concludes that OMBJT's strategy yielded a very useful finding that could not have been revealed by GPUB's strategy.

Effects of the Differing Effect-Size Inclusion Strategies on Findings

Rows 1–5 of Table 1 describe consequences of the two studies' differing effect-size inclusion strategies. Row 1 shows the overall average ICC obtained by OMBJT (with approximate 95% confidence interval), which was $\bar{r} = .132$ ($\pm .04$). Row 2 gives the average correlation for a subset of 100 of these ICCs that appeared identically in the two meta-analyses, $\bar{r} = .259$ ($\pm .06$). Row 3 has 103 ICCs that appeared only in OMBJT—either coming from their 19 added reports or being additional ICCs that either were not available to GPUB or were not used in the same form by GPUB; those 103 ICCs had a smaller mean effect size ($\bar{r} = .096 \pm .06$) than the ones that were included identically in both meta-analyses.

Table 1's Row 4 most directly shows how the differing inclusion strategies affected overall estimated effect sizes. Row 4 includes the approximate half of Row 3's 103 ICCs that were not included or would not have been included by GPUB because they did not fit with GPUB's inclusion policy; that is, they involved neither expected attitude–behavior or stereotype–judgment relationships nor author-expected validity correlations. This subset of 52 ICCs indeed revealed near-null effect sizes ($\bar{r} = .025 \pm .09$). Row 5 aggregates a small complementary set of eight ICCs that GPUB (but not OMBJT) judged to qualify for inclusion ($\bar{r} = .237 \pm .10$). Of these, seven had not been included in OMBJT by their stated policy that excluded "studies of bias against religious groups, obese persons, and older persons" (p. 177); the eighth was an unpublished study of White Australians' bias against Asians (Powell & Williams, 2000) that presumably should have been included by OMBJT.

Effects of Different Sample-Identification Procedures

Table 1's Rows 6 and 7 describe the two meta-analyses' treatments of findings that were reported as 14 ICCs (8 independent samples) by GPUB, but as 38 ICCs (14 independent samples) by OMBJT. OMBJT's larger number of samples resulted partly from their requesting raw data to compute ICCs that authors had not reported and partly from subdividing samples used by GPUB into subsamples for which separate ICCs could be computed. In regard to the latter strategy, if samples are subdivided by randomly assigning cases of a larger sample into two or more subsamples, then the sample splitting should (on average) have no directional effect on magnitudes of aggregate outcomes. However, if samples are split by values on a variable that is similarly correlated with both variables in the correlations being aggregated, the splitting strategy will necessarily reduce the estimated aggregate correlation

[2] OMBJT also commented on the accuracy of some individual effect sizes in GPUB's analyses; indeed, some of those effect sizes were identified as inaccurate in post-GPUB publications. These are not discussed here because, even if all such corrections suggested by OMBJT were treated as valid, their cumulative impact on GPUB's meta-analytic conclusions would be small relative to the effects of differences in inclusion criteria, as described in this section.

[3] Analysis of OMBJT's data set as posted online at http://dx.doi.org/10.1037/a0032734.supp showed that the average correlation for these 33 ICCs for which criterion measures concerned judgments or behavior toward White persons was $\bar{r} = -.020$ with a 95% confidence interval of $\pm .06$.

Table 1
Analysis of Subsets of Effect Sizes (ICCs) for Studies Predicting Discrimination-Relevant Behavior and Judgments

	Number of ICCs	Number of independent samples	Summed weights[a]	Weighted average of independent sample ICCs[b]	Random effects variance (*SD*)	95% CI of weighted average ICC
1. All ICCs in OMBJT	293[c]	98[d]	5,436	**.132**	.110	[.093, .171]
2. ICCs appearing identically in OMBJT and GPUB	100	32	1,085	**.259**	.038	[.201, .316]
3. ICCs in OMBJT but not in GPUB	103	31	3,177	**.096**	.112	[.034, .157]
4. ICCs included by OMBJT and not by GPUB	52	21	489	**.025**	.000	[−.064, .113]
5. ICCs included by GPUB, and not by OMBJT	8	6	316	**.237**	.018	[.130, .339]
6. OMBJT's ICCs in form different from parallels in GPUB	38	14	685	**.070**	.146	[−.042, .180]
7. GPUB's parallels to Row 6's ICCs	14	8	649	**.143**	.038	[.061, .223]
8. OMBJT's ICCs, omitting ones that would have been excluded by GPUB	195	60	4,158	**.177**	.116	[.129, .224]
9. Same as Row 8, limited to Black–White race ICCs	131	43	2,801	**.204**	.131	[.142, .264]

Note. To distinguish them from other numbers in the table, the aggregate ICC values are shown in bold font. This table is based on the full data set of OMBJT and the two (of nine) domains of the GPUB data set to which OMBJT confined their attention. Row 5 includes data from five reports included in GPUB that OMBJT excluded. CI = confidence interval.
[a] Summed weights are the sums of weights for all ICCs regardless of how many of these ICCs were obtained from the same subject sample (as in the analysis method of OMBJT). [b] All aggregated effect sizes and 95% CIs are based on random effects analyses of independent samples to permit comparison with results obtained for the same studies by GPUB. [c] The 293 ICCs in this table for OMBJT are 5 fewer than reported in OMBJT's Table 1. These five were omitted because OMBJT concluded that they had been reported incorrectly in original articles relied on by GPUB, rendering it inappropriate to include them in these comparative analyses. [d] The 98 independent samples for OMBJT are more than the 87 described in their Table 1. However, the data supplement that the authors of OMBJT made available for use in this report identified 98 distinct samples. This discrepancy is largely inconsequential in the present context because OMBJT did not report an independent-sample analysis.

magnitude.[4] Comparison of Rows 6 and 7 of Table 1 shows that there was such a reduction associated with OMBJT's sample splitting. Row 7 shows that, for GPUB's 14 ICCs with average sample size = 83.4 and aggregate effect size of $\bar{r}$ = .143 ± .08, OMBJT created 38 ICCs with average sample size = 64.6 and an aggregate smaller effect size of $\bar{r}$ = .070 ± .11.

Analysis of the Two Data Sets Using the Same Inclusion Criteria

Row 8 of Table 1 estimates the mean ICC effect size that OMBJT would have obtained had they used GPUB's inclusion criteria. It achieves this by dropping OMBJT's ICCs that did not meet GPUB's inclusion criteria—mostly those described in Rows 4 and 6 of Table 1.[5] The aggregate correlation for the remaining 195 ICCs was $\bar{r}$ = .177 ± .05, a value similar to GPUB's overall aggregate effect size ($\bar{r}$ = .220 ± .05) for their combined race and "other intergroup" categories. Although OMBJT's effect was smaller, the two aggregate effect sizes are not significantly different—they are within each other's 95% confidence intervals.

Row 9 of Table 1 reduces Row 8's analysis from 195 of OMBJT's ICCs to 131 by limiting the sample to those for which the predictor was a Black–White race (attitude or stereotype) IAT. The aggregate effect size for these 131 ICCs was $\bar{r}$ = .204 ± .06, which is very close to GPUB's published aggregate effect size estimate of $\bar{r}$ = .236 (± .06) for Black–White race IATs. Compared with OMBJT's full data set (Row 1 of Table 1), this analysis dropped (a) ICCs not involving Black–White race; (b) ICCs involving race, but using criterion measures that were outside GPUB's inclusion policies; and (c) ICCs involving race for which OMBJT had computed their ICCs from subsamples of the samples used by GPUB.

Different Strategies for Testing Moderators of the IAT's Predictive Validity

GPUB identified four significant moderators of the IAT's predictive validity: (a) *correspondence* (operationalized as similarity between verbal descriptions of an IAT measure and the criterion measure with which it was correlated); (b) *social sensitivity* ("the extent to which self-reporting the construct assessed by the measure might activate concerns about the impression that the response would make on others"; GPUB, p. 20); (c) *complementarity* ("the extent to which liking one of the two IAT target categories in a measure implied disliking the other"; GPUB, p. 21), and (d) *implicit–explicit* correlation (operationalized as magnitude of correlation between the IAT measure and parallel self-report measure[s]).

[4] The logic here is straightforward: The numerator of a partial correlation subtracts (from a raw bivariate correlation) the product of the correlations of the (partialed) variable with each of the two variables in the bivariate correlation. When those two correlations with the third variable have the same sign, that product will be numerically positive; therefore, the partial correlation will necessarily be smaller than the raw bivariate correlation. Correlations between two variables that are computed after a sample has been split by values on a third variable show effects similar to partialing by the third variable. When the correlations of the splitting variable with the two correlated variables have the same sign, the splitting will result in averaged correlations of the split samples being reduced relative to the intact original sample. Such effects of disaggregation on magnitudes of correlations are relatives of Simpson's Paradox (Simpson, 1951).

[5] Another set of ICCs dropped from this analysis were ones for which the predicted criteria were facial action coding measures, which have never been demonstrated to relate reliably to discriminatory judgment or behavior.

OMBJT used a recently introduced meta-analysis method (Hedges, Tipton, & Johnson, 2010), in part because of that method's ability to test moderation effects (Oswald et al., p. 177: "this approach allows us to assign criteria to different moderator categories and still model effect-size dependencies"). OMBJT did not test for the four moderators identified by GPUB, but their analyses did reveal two other substantial moderators of ICC magnitudes.

In an analysis that they limited to ICCs involving Black–White race IATs (presented in their Table 3), OMBJT found these ICCs to be larger for criterion measures that assessed differences between behavior toward White and Black persons ($\bar{r} = .22 \pm .12$) or behavior toward only Black persons ($\bar{r} = .15 \pm .085$) than for ICCs involving judgments and behavior toward only White persons ($\bar{r} = -.01 \pm .06$). As mentioned previously, OMBJT's inclusion of this last group (33 ICCs) in computing their overall aggregate predictive validity estimate explained a substantial portion of the difference between their estimate and GPUB's.

The second moderator identified by OMBJT compared prediction by IAT attitude measures with prediction by IAT stereotype measures. Their Table 2 reported substantially stronger ICCs for IAT attitude measures ($\bar{r} = .16 \pm .05$) than for stereotype measures ($\bar{r} = .03 \pm .11$).

Discussion

This article undertook a comparative reexamination of two meta-analytic studies of correlations involving IAT measures—by Greenwald et al. (2009; GPUB) and by Oswald et al. (2013; OMBJT). The comparison was limited to studies of intergroup discrimination, which was a subset of the domain covered by GPUB and was the entire focus of OMBJT. GPUB's meta-analysis set out to examine predictive validity whereas OMBJT's meta-analysis investigated correlations of IAT measures with measures indicative of discrimination, regardless of whether predictive relations were expected. Both meta-analyses found unequivocally significant average positive correlations of IAT measures with criterion measures, although OMBJT's average correlation ($\bar{r} = .141 \pm .045$) was smaller than GPUB's ($\bar{r} = .216 \pm .04$).[6]

OMBJT's authors interpreted their smaller estimate as justifying the conclusion that IAT measures were "poor predictors" (Oswald et al., pp. 171, 182, 183). The present conclusion is that OMBJT's finding of a smaller aggregate effect was due primarily to the difference in purpose of their meta-analysis, which led them to include a substantial minority of effect sizes for which no correlations were theoretically expected. The next section of this discussion describes the two studies' justifications for their differing effect-size inclusion policies, concluding that both strategies were justifiable. The remainder of the discussion describes statistical principles that should be taken into account in evaluating the consequential validity (Messick, 1995) of the two studies' findings.

Both Inclusion Strategies Were Justifiable, Even if Not Equally Suited to Assessing Predictive Validity

GPUB included effect sizes that assessed relations of IAT attitude measures to attitude-consistent behaviors, IAT stereotype measures to stereotype-consistent judgments, and other relationships for which the authors asserted that they expected positive correlations. OMBJT's data inclusion policy was guided by an explicit policy of including "any study for which an [ICC] could be computed where the criterion arguably measured some form of discrimination" (p. 177). Although OMBJT's strategy led to inclusion of ICCs that were excluded by GPUB's strategy, their policy nevertheless is (a) a reasonable one that (b) fits with the well-known methodological strategy of assessing *discriminant validity*. Just as one wishes to know that predictor measures successfully correlate with measures of constructs to which they are conceptually related (*convergent validity*) it is also desirable to know that predictors do not correlate with constructs to which they have no conceptual connection. By limiting their focus to expected predictive validity correlations (a convergent validity strategy) GPUB did not aim to assess discriminant validity.

OMBJT did not explicitly state their discriminant validity findings as conclusions, nor did they connect their discriminant validity findings to the larger body of IAT literature. To take that next step, the present authors examined the 33 "White target" findings in OMBJT's meta-analysis of Black–White race attitude and stereotype IATs. Twenty of these 33 had not been mentioned in the papers and articles in OMBJT's meta-analysis but either were obtained by OMBJT's authors requesting them from authors or were reported by the original study's authors in a publication supplement. Of the 13 effect sizes that were described in original reports, none had been interpreted by authors as having the potential to assess the IAT's predictive validity, and a few were explicitly described as having been obtained as comparison or "control" observations that were not expected to show positive correlations (e.g., Richeson et al., 2003). For example, the correlation between a race attitude IAT and subjects' readiness to judge the face of a White person as showing a friendly expression was not reported by Hugenberg and Bodenhausen (2003) because the authors did not expect it to assess predictive validity.[7] In the 2 (of 33) cases in which predictions of judgments or behavior toward White targets were statistically significant, the studies' authors did not interpret those significant correlations as indicating predictive validity of the IAT (Heider & Skowronski, 2007; Stanley, Sokol-Hessner, Banaji, & Phelps, 2011). Apparently, researchers typically do not expect Black–White race IAT measures to show significant correlation with measures of judgments or behavior toward White persons. Remarkably, that lack of expectation has no strong theoretical basis. Black–White race IAT attitude measures depend as much on variation in evaluations of Whites as on variation in evaluations of Blacks, both of which might be assumed to be involved in race discrimination. Further, theoretical treatments of intergroup attitudes and behavior sometimes emphasize that in-group favorability can be an important contributor to discriminatory outcomes (e.g., Brewer, 1999; Gaertner et al., 1997; Greenwald & Pettigrew, 2014). In this context, OMBJT's finding that "White target" correlations show near-zero correlations

[6] The overall predictive validity estimate for the nine domains in GPUB was $\bar{r} = .274$. The $\bar{r} = .216$ figure given in the text of this paragraph includes just the two of GPUB's nine domains that corresponded to the scope of OMBJT (i.e., White–Black race and "other intergroup" studies). The $\bar{r} = .141$ for OMBJT is the result given as $\bar{r} = .14$ in their Table 1 (see Footnote 1).

[7] Hugenberg and Bodenhausen (2003) wrote: "We hypothesized that high-prejudice European Americans would take longer than their low-prejudice counterparts to respond to Black (but not White) faces changing from hostile to friendly expressions." (pp. 641–642)

with measures of discrimination suggests the need for further development of theory.

Importance of Effect Size in Understanding Consequential Validity

OMBJT characterized their average correlation finding for IAT measures (which they estimated as $\bar{r} = .148$, in the domain of intergroup behavior) as indicating that the IAT was a "poor" predictor (pp. 171, 182, 183). This section's analysis reaches a very different conclusion by applying well-established statistical reasoning to understand the societal consequences of small-to-moderate correlational effect sizes. The first step of this analysis shows that OMBJT's and GPUB's meta-analytic findings had very similar implications for the average percentage of criterion-measure variance explained by IAT measures. The second step explains how statistically small effects can have societally important effects under two conditions—if they apply to many people or if they apply repeatedly to the same person. In combination, the two steps of this analysis indicate how conventionally small (and even subsmall) effect sizes can have substantial societal significance in ways that Messick (1995) characterized in terms of *consequential validity*. Messick defined consequential validity as the aspect of construct validity that "appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (p. 745).

Percentage of Criterion Measure Variance Explained by IAT Measures

For Black–White race attitude and stereotype IATs, GPUB estimated a mean ICC of $\bar{r} = .236$ with no significant heterogeneity of effect size (Greenwald et al., 2009, Table 3). The estimated percentage of criterion variance explained by this result is $r^2 = .236^2 = .056$ (5.6% of variance). OMBJT estimated a smaller mean effect-size ICC of $\bar{r} = .148$. That smaller estimate would appear to imply a smaller percentage ($.148^2 = .022$, or 2.2%) of criterion variance being explained. However, the 2.2% estimate presumes that the expected effect size is fixed across studies. Counter to that presumption, OMBJT's analysis reported substantial random-effects variability for their average effect size estimate, corresponding to a standard deviation of .187 (see their Table 1, 8th data row, *tau* value of .19). For OMBJT's effect size, an expectation for percentage of variance explained could be computed by creating a large simulated distribution of expected true effect sizes based on their mean effect size ($r = .148$) and its random-effects variability ($SD = .187$) and giving a negative sign to squares of negative predictive validities before averaging. The result was an estimated mean percentage of variance explained of 4.4%, which is close to the 5.6% value implied by GPUB's (fixed) estimate of $\bar{r} = .236$.

Societal Significance of Discrimination Predictable From IAT Measures

Among the settings in which IAT measures can be used to predict discrimination are personnel decisions (hiring, performance evaluation, salary, promotion), law enforcement decisions (stops and searches of drivers, pedestrians, or travelers), criminal justice decisions (jury and bench verdicts, sentencing, bail setting, parole, inmate discipline), educational decisions (admissions, grading, disciplinary actions, suspensions), and health-care decisions (triage, treatment authorization, prescription). In all of these settings, IAT measures and other available predictors may be used (a) to identify persons especially prone to committing discrimination and (b) to understand system-level discrimination.

Identifying likely perpetrators of discrimination. IAT measures have two properties that render them problematic to use to classify persons as likely to engage in discrimination. Those two properties are modest test–retest reliability (for the IAT, typically between $r = .5$ and $r = .6$; cf., Nosek et al., 2007) and small to moderate predictive validity effect sizes. Therefore, attempts to diagnostically use such measures for individuals risk undesirably high rates of erroneous classifications.[8] These problems of limited test–retest reliability and small effect sizes are maximal when the sample consists of a single person (i.e., for individual diagnostic use), but they diminish substantially as sample size increases. Therefore, limited reliability and small to moderate effect sizes are not problematic in diagnosing system-level discrimination, for which analyses often involve large samples.

Small effect sizes comprise significant discrimination. For most of the time since the passage of the United States' civil rights laws in the 1960s, U.S. courts have used a statistical criterion of discrimination that translates to correlational effect sizes that are often smaller than $r = .10$. This criterion is the "four-fifths rule," which tests whether a *protected class* (identified by race, color, religion, national origin, gender, or disability status) has been treated in discriminatory fashion. A protected class's members receiving some favorable outcome less than 80% as often as a comparison class can be treated by courts as indicating an "adverse impact" that merits consideration as illegal discrimination (U.S. Equal Employment Opportunity Commission, 1978, §1607.4.D).

Translation of the four-fifths rule's 80% criterion into an effect size such as a correlation coefficient requires assumptions about (a) *base rate*, the overall percentage receiving the favorable treatment, and (b) *class proportion*, the protected class's size in relation to the better-rewarded group. When base rate and class proportion are 50% (i.e., half of the population receives the favorable treatment and the protected class is half of the total population), the four-fifths rule translates to a correlation of $r = .111$.[9]

[8] This caution notwithstanding, there have been proposals to use IAT measures to characterize individual respondents as sexually deviant (e.g, Gray, Brown, MacCulloch, Smith, & Snowden, 2005; Schmidt, Mokros, & Banse, 2013) or as lying (Agosti & Sartori, 2013). These particular individual-diagnostic uses may have better prospects than others because they seek to identify persons whose IAT scores occur infrequently in natural populations.

[9] The *r* values given in these three paragraphs concerning the four-fifths rule are all phi coefficients, which are appropriate for correlations between two dichotomous variables. Translation of the four-fifths rule varies with base rate and class proportion. A smaller base rate results in a smaller phi coefficient. For example, with a 30% base rate, the four-fifths rule (26.7% vs. 33.3%) translates to $r = .073$, and with a 10% base rate (8.9% vs. 11.1%) the four-fifths rule translates to $r = .037$. All of these correlations are reduced if the protected class proportion is (as frequently occurs) smaller than 50% of the total population. With a base rate of 50%, if the class proportion is 30%, then the correlation implied by the four-fifths rule drops from $r = .111$ to $r = .102$, and it drops further to $r = .067$ if the class proportion is only 10% of the total.

In two experiments in which IAT measures were used to predict hiring discrimination against Arab-Muslim applicants by Swedish hiring managers, effect sizes of the ethnicity effect on discrimination were $r = .113$ and $r = .181$ (Rooth, 2010); they were somewhat smaller in a subsequent study that predicted hiring discrimination against obese applicants ($r = .065$ for male applicants and $r = .080$ for female applicants; Agerström, & Rooth, 2011). The relative risk ratios in the two Swedish-Arab studies were 67% and 59%, qualifying as discrimination by the standard of the four-fifths rule; those for the obesity study were slightly outside of the four-fifths rule's standard (83% and 81%).

Small effect sizes predict substantial discrimination in biases affecting many people. As a hypothetical example, assume that a race IAT measure has been administered to the officers in a large city police department, and that this IAT measure is found to correlate with a measure of issuing citations more frequently to Black than to White drivers or pedestrians (profiling). To estimate the magnitude of variation in profiling explained by that correlation, it is necessary to have an estimate of variability in police profiling behavior. The estimate of variability used in this analysis came from a published study of profiling in New York City (Office of the Attorney General, 1999), which reported that, across 76 precincts, police stopped an average of 38.2% ($SD = 38.4\%$) more of each precinct's Black population than of its White population. Using OMBJT's $r = .148$ value as the IAT–profiling correlation generates the expectation that, if all police officers were at 1 *SD* below the IAT mean, the city-wide Black–White difference in stops would be reduced by 9,976 per year (5.7% of total number of stops) relative to the situation if all police officers were at 1 *SD* above the mean. Use of GPUB's larger estimate of $r = .236$ increases this estimate to 15,908 (9.1% of city-wide total stops).

Cumulative impact: Biases affecting the same persons repeatedly. Small effects can produce substantial discriminatory impact also by cumulating over repeated occurrences to the same person. Such repetitions can occur in employment settings (in response to multiple job applications or in periodic performance evaluations in the same job), in educational settings (evaluations of tests and homework by the same student), in health-care settings (repeated patient contacts with medical personnel in successive clinic or hospital visits), and in law enforcement (on city streets and highways).

Many research studies provide evidence for discrimination in situations that allow for repeated impacts on the same person. Audit experiments in field settings typically reveal lower rates of callbacks—invitations to appear for an interview after a job application—for Black and Hispanic than for White applicants (Bendick, 2004; Bertrand & Mullainathan, 2004; Pager, 2003). Experimental studies of evaluations given to work done by men and women often reveal a relative underevaluation of women's achievements (e.g., Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Swim, Borgida, Maruyama, & Myers, 1989). Widespread racial and socioeconomic disparities in health-care interactions have been documented in the Institute of Medicine's (2002) book-length report. An experimental study of responses to e-mailed inquiries from college students found that women and minorities received fewer replies than did White males (Milkman, Akinola, & Chugh, 2012). Studies of school discipline and suspensions have found that Black and Hispanic students are more frequently disciplined in ways likely to result in school dropout (Carter, Fine, & Russell, 2014).

Appraising the cumulative impact of repeated experiences of discrimination is straightforward. With the simplifying assumption that each repetition of an adverse event has the same probabilistic effect, the mathematical description of expected cumulative adverse impact is

$$\text{Impact} = 1 - p^{\text{k}}$$

where p is the probability of a successful (nonadverse) outcome and k is the number of repeated occurrences of the event. The "Impact" estimate computed by the equation is the expected proportion of persons not achieving the successful outcome. If, as a result of discrimination, two groups have different probabilities of successful outcomes, then the discriminatory impact is

$$p_m^k - p_p^k$$

where p_{m} is the probability of successful outcomes to the majority group, and p_{p} is the probability of successful outcomes to the protected class. If p_{m} is a very high value (.99) and p_{k} is only slightly smaller (.98), then with 5, 10, 15, 20, or 25 repetitions, discriminatory impacts will disadvantage, respectively, 4.7%, 8.7%, 12.1%, 15.0%, and 17.4% of the protected class. Slightly enlarging the separation between p_{m} (.995) and p_{k} (.975) approximately doubles these estimates of discriminatory disadvantage. The correlational effect size of the .99 versus .98 difference is $r = .041$; for the larger difference (.995 vs. .975) it is $r = .082$.

Not a new observation. Rosenthal (1990) drew attention to the outcome of a large randomized clinical trial of the effect of aspirin in preventing heart attacks among male physicians. The trial was terminated early because data analysis had revealed an unexpected effect for which the correlational effect size was the sub-small value of $r = .035$. This was "a significant ($p < .00001$) reduction [from 2.16% to 1.27%] in the risk of total myocardial infarction [heart attack] among those in the aspirin group" (Steering Committee of the Physicians' Health Study Research Group, 1989). Applying the study's estimated risk reduction of 44% to the 2010 U.S. Census estimate of approximately 46 million male U.S. residents 50 or older, regular small doses of aspirin should prevent approximately 420,000 heart attacks during a 5-year period.[10]

Continuing the spirit of Rosenthal's analyses, Figure 1 describes theoretical expectations for predictive efficacy of correlations ranging from $r = .05$ to $r = .30$ when used to predict above-threshold performance on a criterion measure. The computations of Figure 1 assume an interest in predicting who will exceed thresholds ranging from 2% to 98% of the population on a criterion measure, further assuming that all of those exceeding a given percentile threshold on a quantitative predictor will be predicted to exceed that same threshold on a dichotomous criterion (i.e., above vs. below threshold). The plot displays the accuracy gain from those predictions relative to chance prediction, which can be understood as prediction by a variable that has a correlation of $r = .00$ with the criterion measure. For example, with a 50th percentile threshold, random prediction such as a coin toss should be 50% correct. For a predictor that correlates with a dichotomous criterion at point biserial r [r_{pb}] = .30, 62.3% of those above the 50th

[10] The U.S. Census estimate of male population by age was accessed on December 15, 2013 at http://www.census.gov/compendia/statab/cats/population/estimates_and_projections_by_age_sex_raceethnicity.html

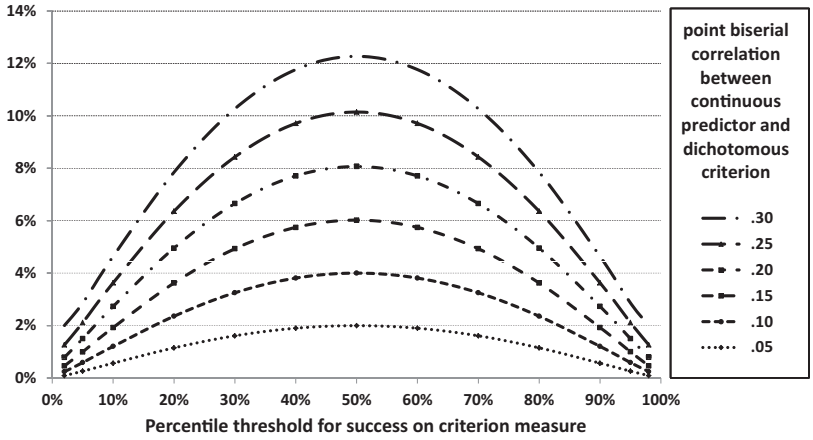

Figure 1. Gain in percentage of cases predicted correctly by predictors with small to moderate correlational effect sizes compared with a predictor having zero correlation with the criterion. The "threshold" is a percentile treated as the indicator of "success" on the criterion measure. The prediction is in the form of predicting that all (and only) those who exceed the threshold percentile on the predictor will pass that threshold on the criterion.

percentile on the predictor should exceed the 50th percentile on the criterion, affording an accuracy increase of 12.3% (the highest plotted value in Figure 1) relative to prediction in ignorance. The noticeable decline in accuracy gain as the threshold is either lowered or raised is due to the increase in chance accuracy. For example, with the threshold at the 90th percentile and a predictor that has zero correlation with the criterion, 82% ($= .10^2 + .90^2$) of predictions (i.e., that a random 10% of a sample will exceed threshold and the remaining 90% will fall short) should be correct by chance.[11]

For small to moderate effects, ranging from Cohen's $d = 0.1$ to $d = 0.5$, Figure 2 shows effects of interventions aimed at improving performance on a criterion measure. The computations of Figure 2 assume that the intervention is targeted at a random sample of a reference population. For each intervention effect size, the plotted values are the fraction of the intervention sample expected to exceed the *x*-axis percentile minus the comparable fraction for the reference population (i.e., those not receiving the intervention). An example: If (a) the reference population is an electorate evenly divided between Candidates A and B and (b) the 50th percentile serves as the boundary separating those expected to vote for A from those expected to vote for B, then a pro-B intervention with $d = 0.2$ (Cohen's [1977] conventional "small" effect size) will increase B's vote by 7.9%, such that the intervention target group should show a 57.9:42.1 ratio favoring B. If that intervention can be administered to 10% of the electorate, then Candidate B should have a possibly comfortable victory margin of 1.58%.

Growing recognition of the significant cumulative impact of very small acts of discrimination is signaled by the invention of three terms to label it—*microinequities* (by Mary Rowe, when ombudsperson at Massachusetts Institute of Technology in 1973), *micro acts of discrimination* (Reskin, 2002), and *microaggressions* (Sue et al., 2007). Virginia Valian (1998) also strongly advanced the thesis that minor acts of discrimination have significant cumulative impact on women's professional careers. Earlier than all of these, sociologist Robert K. Merton (1968) described the *Matthew effect* as a "cumulative advantage [that] operates in many systems of social stratification to produce the same result: the rich get richer at a rate that makes the poor become relatively poorer."

Conclusions

This article drew two conclusions from analysis of the apparent disagreement between meta-analytic conclusions of Greenwald et al. (2009; GPUB) and Oswald et al. (2013; OMBJT). First, differences in the two meta-analyses' published conclusions were due to differences in the methods they used. Second, both meta-analyses estimated aggregate correlational effect sizes that are large enough to justify concluding that IAT measures predict societally important discrimination.

The main method difference between the two studies was in their respective policies for including effect sizes. GPUB limited their meta-analysis to effect sizes for which there was reason to expect nontrivial predictive validity correlations of IAT measures with criterion measures. OMBJT included numerous additional effect sizes that lacked a basis either in existing theory or in author-provided rationale for expecting positive correlations. GPUB explicitly described their article as a "meta-analysis of predictive validity," whereas OMBJT did not describe a goal of assessing predictive validity—they instead described their study as a "meta-analysis of IAT criterion studies." This important strategy difference, with its concomitant difference in policies for including effect sizes, explains most of the difference between the average effect sizes that the two meta-analyses estimated.

A second important difference was in the two studies' judgments of societal significance of their observed effect sizes. GPUB

[11] Rosenthal (1990; Rosenthal & Rubin, 1982) used the phi coefficient rather than the point biserial that is used in this article's Figure 1. The phi coefficient is appropriate when the predictor and criterion are dichotomous. The phi coefficient also affords larger estimates for accuracy gain than does the point biserial. For example, with $\phi = .30$, the accuracy gain when predictor and criterion thresholds are at the 50th percentile is 15%, which is greater than the 12.3% gain value shown in Figure 1.

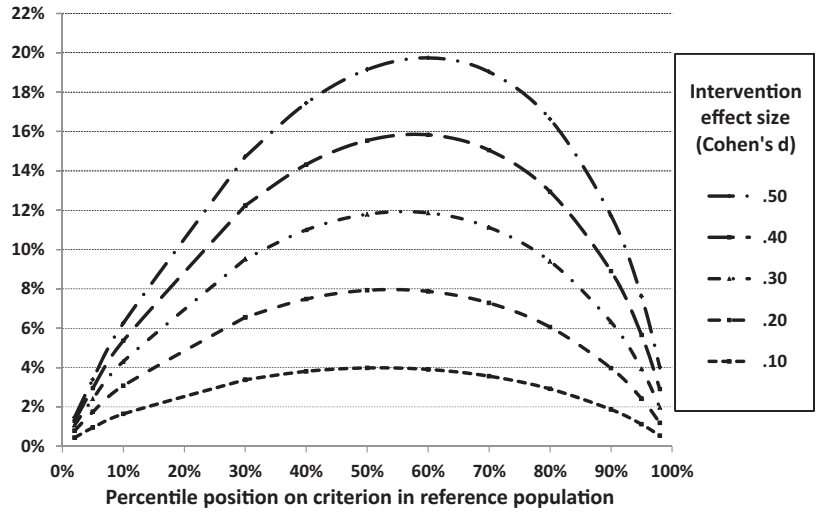

Figure 2. Effects of small to moderate effect size interventions (in units of Cohen's *d*) on fraction of an intervention sample surpassing a comparable fraction of a reference (no-intervention control) population on a criterion variable. Values are plotted as a function of percentile value treated as a "passing" score on the criterion measure.

did not comment on societal significance in their article, whereas OMBJT concluded that IAT measures show "poor prediction of racial and ethnic discrimination" (pp. 171, 183) and provide "little insight into who will discriminate against whom" (p. 188). OMBJT's conclusion did not take into account that small effect sizes affecting many people or affecting individual people repeatedly can have great societal significance.

Differences between the conclusions of the two meta-analyses notwithstanding, two important empirical findings were supported by both. First, both studies agreed that, when considering only findings for which there is theoretical reason to expect positive correlations, the predictive validity of Black–White race IATs is approximately $r = .20$. Second, even using the two meta-analyses' published aggregate estimated effect sizes, the two agreed in expecting that more than 4% of variance in discrimination-relevant criterion measures is predicted by Black–White race IAT measures. This level of correlational predictive validity of IAT measures represents potential for discriminatory impacts with very substantial societal significance.

References

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96,* 790–805. http://dx.doi.org/10.1037/a0021594

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology, 4,* 519. http://dx.doi.org/10.3389/fpsyg.2013.00519

Ajzen, I., & Fishbein, M. (1977). Attitude– behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84,* 888–918. http://dx.doi.org/10.1037/0033-2909.84.5.888

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91,* 652–661. http://dx.doi.org/10.1037/0022-3514.91.4.652

Bendick, M. (2004, June). *Using paired-comparison testing to develop a social psychology of civil rights*. Paper presented at the biennial conference of the Society for the Psychological Study of Social Issues, Washington, DC.

Bertrand, M., & Mullainathan, S. (2004). *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination*. Chicago, IL: University of Chicago Business School.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues, 55,* 429–444. http://dx.doi.org/10.1111/0022-4537.00126

Carter, P., Fine, M., & Russell, S. (2014). *Discipline disparities series: Overview*. Bloomington, IN: Center for Evaluation and Education Policy.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). New York: Academic Press.

Gaertner, S. L., Dovidio, J. F., Banker, B. S., Rust, M. C., Nier, J. A., Mottola, G. R., & Ward, C. M. (1997). Does racism necessarily mean anti-Blackness? Aversive racism and pro-Whiteness. In M. Fine, L. Powell, L. Weis, & M. Wong (Eds.), *Off white: Readings on race, power, and society* (pp. 167–178). London, United Kingdom: Routledge.

Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology, 114,* 304–308. http://dx.doi.org/10.1037/0021-843X.114.2.304

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48,* 85–93.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216. http://dx.doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69,* 669–684. http://dx.doi.org/10.1037/a0036056

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97,* 17–41. http://dx.doi.org/10.1037/a0015575

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. http://dx.doi.org/10.1002/jrsm.5

Heider, J. D., & Skowronski, J. J. (2007). Improving the predictive validity of the Implicit Association Test. *North American Journal of Psychology, 9,* 53–76.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14,* 640–643. http://dx.doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Institute of Medicine. (2002). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, DC: National Academy of Sciences.

Kraus, S. J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and Social Psychology Bulletin, 21,* 58–75. http://dx.doi.org/10.1177/0146167295211007

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159,* 56–63. http://dx.doi.org/10.1126/science.159.3810.56

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science, 23,* 710–717. http://dx.doi.org/10.1177/0956797611434539

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 16474–16479.

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

Office of the Attorney General. (1999). The New York City Police Department's "stop & frisk" practices: A report to the people of the State of New York. Retrieved from www.oag.state.ny.us/bureaus/civil_rights/pdfs/stp_frsk.pdf

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105,* 171–192. http://dx.doi.org/10.1037/a0032734

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108,* 937–975. http://dx.doi.org/10.1086/374403

Phelps, E. A., O'Conner, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12,* 729–738. http://dx.doi.org/10.1162/089892900562552

Powell, A. J., & Williams, K. D. (2000, April). *The role of racial attitudes and experience in cross-racial identification among Asian, Caucasian and Eurasian Australians*. Paper presented at the 5th annual meeting of the Society of Australasian Social Psychologists, Fremantle, Western Australia.

Reskin, B. F. (2002). Rethinking employment discrimination and its remedies. In M. Guillen, R. Collins, P. England, & M. Meyer (Eds.), *The new economic sociology: Developments in an emerging field* (pp. 218–244). New York, NY: Russell Sage Foundation.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI examination of the impact of interracial contact on executive function. *Nature Neuroscience, 6,* 1323–1328. http://dx.doi.org/10.1038/nn1156

Rooth, D.-O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17,* 523–534. http://dx.doi.org/10.1016/j.labeco.2009.04.005

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777. http://dx.doi.org/10.1037/0003-066X.45.6.775

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169. http://dx.doi.org/10.1037/0022-0663.74.2.166

Shelton, J. N., Richeson, J. A., Salvatore, J., & Trawalter, S. (2005). Ironic effects of racial bias during interracial interactions. *Psychological Science, 16,* 397–402.

Schmidt, A. F., Mokros, A., & Banse, R. (2013). Is pedophilic sexual preference continuous? A taxometric analysis based on direct and indirect measures. *Psychological Assessment, 25,* 1146–1153. http://dx.doi.org/10.1037/a0033326

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B. Methodological, 13,* 238–241.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 108,* 7710–7715. http://dx.doi.org/10.1073/pnas.1014345108

Steering Committee of the Physicians' Health Study Research Group. (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. *The New England Journal of Medicine, 321,* 129–135. http://dx.doi.org/10.1056/NEJM198907203210301

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62,* 271–286. http://dx.doi.org/10.1037/0003-066X.62.4.271

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105,* 409–429. http://dx.doi.org/10.1037/0033-2909.105.3.409

U.S. Equal Employment Opportunity Commission. (1978). *Uniform Guidelines on Employee Selection Procedures*. Retrieved from http://www.gpo.gov/fdsys/pkg/CFR-2013-title29-vol4/xml/CFR-2013-title29-vol4-part1607.xml

Valian, V. (1998). *Why so slow? The advancement of women*. Cambridge, MA: MIT Press.

Received January 18, 2014
Revision received September 4, 2014
Accepted September 4, 2014 ■