Article accepted for publication, *American Psychologist*

Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis

Benedek Kurdi
Harvard University
Cambridge, Massachusetts

Allison E. Seitchik
Merrimack College
North Andover, Massachusetts

Jordan R. Axt
University of Virginia
Charlottesville, Virginia

Timothy J. Carroll
Harvard University
Cambridge, Massachusetts

Arpi Karapetyan
Harvard University
Cambridge, Massachusetts

Neela Kaushik
Boston University
Boston, Massachusetts

Diana Tomezsko
Harvard University
Cambridge, Massachusetts

Anthony G. Greenwald
University of Washington
Seattle, Washington

Mahzarin R. Banaji
Harvard University
Cambridge, Massachusetts

Author Note

Benedek Kurdi, Department of Psychology, Harvard University; Allison E. Seitchik, Department of Psychology, Merrimack College; Jordan R. Axt, Department of Psychology, University of Virginia; Timothy J. Carroll, Department of Psychology, Harvard University; Arpi Karapetyan, Graduate School of Education, Harvard University; Neela Kaushik, Department of Computer Science, Boston University; Diana Tomezsko, Harvard Law School, Harvard University; Anthony G. Greenwald, Department of Psychology, University of Washington; Mahzarin R. Banaji, Department of Psychology, Harvard University.

Benedek Kurdi and Allison E. Seitchik contributed equally to this work. Jordan R. Axt, Timothy J. Carroll, Arpi Karapetyan, Neela Kaushik, and Diana Tomezsko are listed alphabetically.

We thank the authors of the studies included in this meta-analysis. Without their significant efforts to share additional effect sizes and data files with us, the estimates reported here would have been considerably less accurate. A full list of these authors as well as the papers for which no data could be obtained can be found in Supplement 1. We also thank Neeha Dhawan, Alex Garinther, and Sarah Vasconcelos for study coding, as well as Abi Cherry, Ellie Cherry, Christina Dias, Catherine Kim, Ruolin Lu, Sarah Ryan, and Jared Valdron for assistance with study coding. The project was supported by a grant from the Edmond J. Safra Center for Ethics at Harvard University to Mahzarin R. Banaji.

This article is based entirely on open data and fully reproducible analyses. The raw data file and analysis scripts as well as all supplementary materials are available for download from the Open Science Framework (https://osf.io/47xw8/).

Correspondence concerning this article should be addressed to Benedek Kurdi, Department of Psychology, Harvard University, Cambridge, MA 02138, email: kurdi@g.harvard.edu, or Allison E. Seitchik, Department of Psychology, Merrimack College, North Andover, MA 01845, email: seitchika@merrimack.edu.

Abstract

Using data from 217 research reports (N = 36,071, compared to 3,471 and 5,433 in previous meta-analyses), this meta-analysis investigated the conceptual and methodological conditions under which Implicit Association Tests (IATs) measuring *attitudes*, *stereotypes*, and *identity* correlate with criterion measures of *intergroup behavior*. We found significant implicit–criterion correlations (ICCs) and explicit–criterion correlations (ECCs), with unique contributions of implicit (β = .14) and explicit measures (β = .11) revealed by structural equation modeling. ICCs were found to be highly heterogeneous, making moderator analyses necessary. Basic study features or conceptual variables did not account for any heterogeneity: Unlike explicit measures, implicit measures predicted for all target groups and types of behavior, and implicit, but not explicit, measures were equally associated with behaviors varying in controllability and conscious awareness. However, ICCs differed greatly by methodological features: Studies with a declared focus on ICCs, standard IATs rather than variants, high-polarity attributes, behaviors measured in a relative (two categories present) rather than absolute manner (single category present), and high implicit–criterion correspondence (k = 13) produced a mean ICC of r = .37. Studies scoring low on these variables (k = 6) produced an ICC of r = .02. Examination of methodological properties—a novelty of this meta-analysis—revealed that most studies were vastly underpowered and analytic strategies regularly ignored measurement error. Recommendations, along with online applications for calculating statistical power and internal consistency (http://www.benedekkurdi.com/#iat), are provided to improve future studies on the implicit–criterion relationship. Open materials are available under https://osf.io/47xw8/.

Keywords: Implicit Association Test, implicit social cognition, intergroup relations, meta-analysis, predictive validity

Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis

In recent decades, the study of social cognition, in particular the study of attitudes and beliefs about social groups, has been dominated by the introduction of indirect measures of mental content (Fazio & Olson, 2003). Unlike measures of explicit cognition that rely on self-report to access mental content, measures of implicit cognition rely on less controllable behaviors, such as response latencies or other responses that bypass conscious awareness. Early use of such measures was to study category structure and semantic associations (Meyer & Schvaneveldt, 1971; Neely, 1976). For instance, it was demonstrated that participants are faster to respond to a target word like *nurse* if it is preceded by a semantically related word such as *doctor*, as opposed to a semantically unrelated word like *bread*. In the 1980s, researchers began using implicit measures to reveal the representation of social categories, and found that White and Black primes facilitated responding to evaluatively or stereotypically consistent stimuli (Devine, 1989; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Gaertner & McLaughlin, 1983).

The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is one such measure, based on the longstanding assumption that the speed and accuracy of responses can serve as useful indicators of underlying mental processes (Luce, 1986). The IAT measures participants' response latencies and accuracy in combined categorizations of category (e.g., *young* vs. *elderly*) and attribute stimuli (e.g., *good* vs. *bad*). The relative strength of association between categories and attributes is inferred from differences in response latencies across two types of trials: Ones in which participants sort stimuli in a congruent manner (e.g., same response for young and good stimuli and same response for elderly and bad stimuli) vs. ones in which they sort stimuli in an incongruent manner (e.g., same response for young and bad stimuli and same response for elderly and good stimuli). For a demo of the IAT, visit the Project Implicit educa-

tional website (http:/implicit.harvard.edu/). The IAT's adaptation as a method is visible in the over 9,500 citations it has received in theoretical and empirical reports of social cognition broadly, including domains such as group perception, person perception, consumer preferences, close relationships, personality, clinical disorders, and political behavior. Together, these studies have provided new insights into the nature of implicit social cognition, by revealing construct validity (Nosek, Greenwald, & Banaji, 2005), dissociation and association between implicit and explicit measures (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005), cultural and geographic variation (Nosek et al., 2007), neural underpinnings (Phelps et al., 2000), developmental comparisons (Dunham, Chen, & Banaji, 2013), short-term malleability (Lai et al., 2014), and change over cultural time (Charlesworth & Banaji, 2018).

In addition to dozens of studies that have established construct validity, a substantial number of papers have examined the relationship between performance on the IAT and behavior in the domain of intergroup relations. That is, investigators have asked to what extent, and under what conditions, individual differences in implicit attitudes, stereotypes, and identity are associated with variation in behavior toward individuals as a function of their social group membership.[1] The interest in this issue stems from at least two sources. First, the relationship between explicit measures of attitudes or beliefs and related behaviors has traditionally been used as an indication of the validity of self-report data (LaPiere, 1934; Wicker, 1969). Second, an association between intergroup cognition and intergroup behavior has implications for societies committed to the ideals of equality and non-discrimination. Application of these results can be seen in

[1] Of the three constructs, attitudes and stereotypes are most commonly investigated in the context of implicit social cognition. Identity was also selected for inclusion because (*a*) the self is central to a host of psychological theories (Swann & Bosson, 2010) and (*b*) there has been a steep increase in the number of studies probing the relationship between implicit identity and behavior, with a single effect size eligible for inclusion prior to 2007, compared to 33 after 2007.

the domains of law, education, business, and healthcare: Implicit measures are often used to understand why group-based inequities remain in spite of vast progress in explicit intergroup attitudes and beliefs, which are often near neutral levels (Charlesworth & Banaji, 2018).

Typical examples of studies on the implicit–criterion relationship include Amodio and Devine (2006), in which seating distance from an African American target was predicted by a race attitude IAT. This study is also typical of the field in that it was conducted in a lab setting. Less typical examples include the following two studies. Rüsch, Todd, Bodenhausen, Olschewski, and Corrigan (2010) correlated a mental illness/physical disability–shameful/proud brief IAT with perceived legitimacy of discrimination against individuals with mental illness in a sample of people with mental illness. Agerström and Rooth (2011) conducted a field study with human resources managers in Sweden in which an obese/normal weight–high performance/low performance stereotype IAT was used to predict hiring discrimination.

Over the past decade, two meta-analyses have been published on the relationship between the IAT and measures of behavior (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Greenwald and colleagues' meta-analysis included studies in all available domains, ranging from intergroup discrimination and consumer preferences to personality and close relationships. Oswald and colleagues conducted a second meta-analysis, restricting their focus to studies involving race and ethnicity. Since then, several significant developments have occurred, making a new meta-analysis, drawing a different perimeter around existing studies, a necessity.

The first development justifying a new meta-analysis emerged from the sheer amount of research on the relationship between the IAT and behavior generated over a short period of time: In the domain of intergroup discrimination, the Greenwald et al. (2009) meta-analysis synthe-

sized data from 3,471 participants and 62 research reports. Less than a decade later, the current project was able to aggregate data from 36,071 participants across 217 research reports. Second, the studies conducted since the publication of the Greenwald et al. meta-analysis have not only grown in number but have diversified in the quality and quantity of variables represented across investigations, including the target groups and types of behaviors examined. For instance, the studies included in previous meta-analyses by Greenwald et al. and Oswald et al. focused primarily on race and ethnicity, either due to availability or due to theoretical considerations. By contrast, the current project covers all domains of intergroup discrimination, including studies on gender, age, sexuality, eating disorders, and other clinical conditions. In addition, since 2007, studies have utilized novel and socially consequential target behaviors, such as resource allocation, physical and mental health, and academic performance. Finally, all research reports reviewed by Greenwald and colleagues relied on work conducted in a laboratory setting, whereas the current project was able to include a substantial number of studies conducted in real-world and online settings, using a considerably more diverse pool of participants and ecologically meaningful measures of behavior. Third, advances in statistical methodology now allow for explicit modeling of dependencies among effect sizes extracted from the same study (Hedges, Tipton, & Johnson, 2010) as well as for the appropriate treatment of measurement error (Westfall & Yarkoni, 2016). Fourth, recent developments in the way psychological science is conducted, along with an increased focus on the transparency of statistical analyses, the replicability of results, and the standards for what constitutes good statistical evidence (e.g., Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) also add to the need for a new review. The recent push toward open science makes it a particularly fitting time to produce a meta-analytic database that is sufficiently rich to be useful for additional investigation, and the preparation of this analy-

sis was guided by that imperative. Fifth, research on implicit social cognition has witnessed higher levels of attention both from the general public and from governmental and commercial entities, making regular reporting of what is known an added responsibility.

No meta-analysis can answer all possible questions of interest. Therefore, following a discussion of overall meta-analytic results, including heterogeneity in effect sizes as well as incremental predictive validity of implicit over explicit measures and *vice versa*, our focus here is twofold. First, we report discoveries about potential areas of improvement in study design and analytic strategies. Second, we discuss three groups of more traditional moderator variables, including (*a*) basic study characteristics, (*b*) conceptual variables, and (*c*) methodological variables with implications for theory.

Methodological shortcomings of the reviewed studies

We view discoveries about shortcomings in study design and data analysis as one of the primary contributions of the present work. As such, in addition to the effects of specific moderators discussed below, we address general issues of incremental predictive validity and measurement error, publication bias, and statistical power. We believe that these methodological considerations are paramount: If valid inferences about the relationship between implicit cognition and behavior are of interest to the field of social cognition and, by extension, to the general public, relevant studies must be designed and conducted in such a way as to allow for such inferences. The methodological shortcomings of current studies make it especially important to focus on this problem.

Basic study characteristics

A first group of moderator variables investigated concerns basic study characteristics such as the *target group of the study* (e.g., race, gender, or sexuality), the *type of criterion behav-*

ior measured (e.g., person perception, resource allocation, and nonverbal measures), as well as the *setting in which the study took place* (lab, online, or real-world setting). These variables stand to provide unique insight due to their novelty; previous meta-analyses simply did not possess the variability needed to test their effects.

Conceptual moderators

The second group of moderators concerns various aspects of theories developed in the domain of implicit social cognition. Given potential conceptual traps, we have chosen to refer to the implicit–explicit distinction in terms of measures rather than in terms of underlying mental systems. However, the outcome of this meta-analysis will have implications for mental systems, especially dual-process theories (Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000) as well as classic theories of the attitude–behavior relationship (Ajzen & Fishbein, 1977; Talaska, Fiske, & Chaiken, 2008).

A primary set of conceptual variables emerged from existing theories about the implicit–explicit distinction. First, it has been argued that the *social sensitivity* of the domain, i.e., the extent to which expressing an attitude or performing a behavior evokes concerns about appearing prejudiced, may influence the strength of the relationship between attitudes and behavior and do so differentially for explicit and implicit measures. Stronger impact of social sensitivity may be expected on explicit–criterion than implicit–criterion correlations. Second, a central element embedded in many definitions of implicit attitudes and stereotypes is lack of *controllability*, i.e., automatic activation of mental content upon encountering a stimulus (Fazio et al., 1986). The notion of implicit social cognition as uncontrollable makes it plausible that implicit measures should be especially highly correlated with behaviors that are themselves expressed relatively automatically (e.g., nonverbal and other spontaneous behaviors). Similarly, a hallmark of many

definitions of implicit social cognition is *lack of conscious awareness*, i.e., the idea that, unlike their explicit counterparts, implicit attitudes and stereotypes are not amenable to introspection (Greenwald & Banaji, 1995). To the extent that this is the case, implicit attitudes may be more closely associated with behavior in studies where the hypothesis lies beyond participants' conscious awareness. Conversely, the opposite relationship may be expected to emerge for explicit measures of cognition: When controllability and conscious awareness are high, participants have the ability to infer the meaning and goals implied by the situation, and may, as a result, respond consistently to the explicit and criterion measures.

Overall, dual-process theories posit a dissociation between implicit–criterion and explicit–criterion correlations, with direct and separable effects of implicit and explicit cognition on behavior. However, in a recent paper, Greenwald and Banaji (2017) offered a possible alternative, noting that "[w]hen people attempt to report on their conscious perceptions and judgments, they do so not based on valid introspection, but by using traces of past (possibly biased) experience to construct (possibly invalid) theories of current data" (p. 868). The new idea here is that implicit cognition may not affect behavior directly; rather, it may do so by reshaping conscious cognition. The unique prediction of this approach concerns the relationship between implicit–explicit and implicit–criterion correlations. Namely, if Greenwald and Banaji's theorizing is accurate, implicit measures should be more highly correlated with behavior when implicit and explicit measures are more highly correlated with each other.

A second set of conceptual variables were derived from theories of explicit cognition and accompanying measures that dominated the field for six decades starting in the 1930s. From the work of Ajzen and Fishbein (1977), we know that cognition and behavior ought to be related to the extent that there is *correspondence* between the two, i.e., if measures of cognition and behav-

ior are driven by the same causal processes. That is, attitudes measuring specific components that are present in the criterion should produce stronger attitude–behavior relationships. The implicit–criterion pairs included in the present meta-analysis differed considerably in terms of their levels of correspondence. In some cases, the relationship was easy to grasp based on theory or past research, such as a proposed association between implicit science identity and enrollment intentions in mathematics classes (Steffens, Jelenec, & Noack, 2010); in other cases, the relationship was tenuous, such as a proposed association between implicit race attitudes and tobacco use (Krieger et al., 2011). Second, from its inception, research on social cognition has focused on the triad of attitudes, stereotypes, and identity. A meta-analysis by Talaska, Fiske, and Chaiken (2008) found higher correlations between explicit *attitudes* and behavior than between explicit *stereotypes* and behavior. The present project can investigate whether this result generalizes to the relationship between measures of implicit cognition and intergroup behavior.

Methodological moderators

First, we used lenient inclusion criteria to provide a conservative estimate of the implicit–criterion correlation. As such, we were able to newly probe the effect of *study focus*, i.e., whether studies that explicitly set out to explore the relationship between implicit social cognition and behavior tend to produce larger effects than studies that incidentally included both measures.

Second, we investigated whether implicit–criterion correlations are modulated by differences in the implicit measure used. There are many variations that the IAT has spawned, each addressing what is viewed as a methodological limitation. Specifically, we tested for differences in the magnitude of the implicit–criterion relationship depending on the *type of implicit measure*, comparing the standard Implicit Association Test (Greenwald et al., 1998) to its variants such as the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006), the single-

category IAT (SC-IAT; Karpinski & Steinman, 2006), the brief IAT (B-IAT; Sriram & Greenwald, 2009), the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001), the personalized IAT (M. A. Olson & Fazio, 2004), and the paper-and-pencil IAT (Sekaquaptewa, Vargas, & von Hippel, 2010). On the one hand, the IAT may be expected to produce larger effect sizes because of its superior internal consistency (Bar-Anan & Nosek, 2014). On the other hand, each new variant of the IAT featured some methodological innovation, which may increase its association with behavioral measures. Moreover, a traditional IAT includes two targets (e.g., *men* vs. *women*) and two attributes (e.g., *pleasant* vs. *unpleasant*). The attribute pairs used in IATs differ along the dimension of *polarity*, i.e., the extent to which the two concepts are polar opposites of each other. High-polarity attributes (e.g., *fat* vs. *thin*) may produce larger effects than low-polarity attributes (e.g., *sad* vs. *angry*) because they tap into a more cohesive network of mental representations.

Finally, the *way in which criterion behaviors are measured* may affect the magnitude of implicit–criterion correlations. Specifically, ICCs may be higher when behaviors are measured in a relative way (i.e., involving both categories of interest) rather than in an absolute way (i.e., involving only one category). For instance, a race attitude IAT may be more highly correlated with the difference between donations to a White vs. Black student group than with donations to a Black student group only and especially to a White student group only. Such a result may emerge because the IAT is an inherently relative measure. Moreover, psychologically, acts of intergroup discrimination may be conceptually relative: They often involve tradeoffs in affect, beliefs, and especially behavior, e.g., between who receives a resource, whether it be a material one like money or a non-material one like helping (Carlsson & Agerström, 2016).

Method

Below we highlight the most important aspects of the literature search, study eligibility and coding, and analytic strategy. All further details of the meta-analytic methods, such as the specific search terms and Boolean operators used, coder training, procedure for including and excluding studies and effect sizes, and moderators coded, are provided in Supplements 2–5.

Literature search

To ensure valid inferences, we strove to include all pertinent studies in the meta-analytic database. As such, considerable effort was devoted to the search and retrieval of potentially eligible research reports, both published and unpublished. The search and retrieval process used three main sources: (*1*) relevant previous meta-analyses (Greenwald et al., 2009; Oswald et al., 2013); (*2*) three online search engines with divergent scopes, including PsycINFO (psychology journals and dissertations), Web of Science (scientific journals from all disciplines including economics, business, and medicine), and HeinOnline Law Journal Library (law journals); and (*3*) an open call for unpublished research reports sent out to the Society of Personality and Social Psychology listserv. We sought to identify eligible research reports that were published or completed prior to February 28, 2016, which served as the cutoff date for the present project. For an overview of the literature search and screening process, a list of excluded research reports and reasons for exclusion, as well as a list of missed studies discovered after completion of this meta-analysis and comparison to missed studies in Oswald et al. (2013), see Supplement 3.

Study eligibility

The initial meta-analysis on the relationship between the IAT and behavior (Greenwald et al., 2009) synthesized effect sizes from a wide range of domains including political behavior, consumer preferences, and close relationships. By contrast, the present work concentrates exclusively on studies in the domain of intergroup relations. Nevertheless, its scope is broader than the

meta-analysis conducted by Oswald et al. (2013), which covered only studies involving race and ethnicity. As such, the present project focused on the relationship between implicit social cognition and intergroup behavior broadly defined, including race and ethnicity, gender, sexual orientation, age, religion, as well as psychiatric and medical conditions. In line with this focus, studies were eligible for inclusion in the meta-analytic database if they met the following three criteria. First, studies had to report an empirical result. Second, studies had to use the IAT (Greenwald et al., 1998) or one of its variants. Third, studies had to contain a measure of behavior toward a group or toward individuals as members of social groups (criterion measure). Additional criteria for inclusion and exclusion are listed in Supplement 3. Some studies also reported an explicit (self-report) measure paralleling the implicit measure; however, this was not a criterion for eligibility.

In determining the criteria for inclusion of studies, authors of meta-analyses must make choices about papers that may need to be discarded as a function of poor instantiation of theoretical variables or other identifiable errors in method. We found several such papers but decided against exclusion so as to offer the most conservative estimate of the effect size possible. For instance, a study that makes the prediction that degree of race bias will affect smoking behavior was included. Likewise, results that do not fit with existing theory but might serve as the basis for developing new theory were included as showing a negative implicit–criterion relationship. For example, contrary to common intuition, Rae, Newheiser, and Olson (2015) revealed that greater racial diversity in a region can be associated with higher levels of race bias. Although theoretical accounts have since been provided for this result, it was still included as a negative effect.

The online search using the three databases mentioned above yielded 3,855 research reports potentially eligible for inclusion, with 64 additional research reports obtained from previous meta-analyses, and four from the listserv request. Each of these close to 4,000 research reports, which constituted approximately half of the entire IAT literature at the time, was individually screened by trained coders (see Supplement 2). As a result of this screening process, 295 research reports were identified that met inclusion criteria. Of these, 183 were found not to include the data necessary to calculate the effect sizes of interest. In all such cases, the corresponding author was contacted via email and asked to calculate effect sizes or share data files that would allow for their calculation. If the data could not immediately be obtained, authors were sent at least two reminder emails. Of the 183 corresponding authors contacted, 134 (i.e., 73 percent) shared effect sizes or data files. As a result, a total of 246 research reports could be included in the meta-analytic database. Of these, most statistical analyses focus on the 217 research reports from which effect sizes could be obtained separately for independent samples (see below).[2]

Coding of effect sizes and moderator variables

Each study, including effect sizes and moderator variables (see below), was coded by two trained coders. Effect sizes were checked by a third and fourth coder to eliminate any remaining errors. In addition, to remove biasing influences, all conceptual moderators involving a relatively high degree of subjectivity (i.e., correspondence, social sensitivity, controllability, and conscious awareness) were coded by an additional coder blind to study results. Blind coding was performed on the basis of a summary containing the study hypotheses, independent samples, procedure, as well as implicit, explicit, and criterion measures (see Supplement 4).

[2] The data file made available on OSF also contains a single summary effect size for each study that collapses across independent samples.

Interrater reliability. Any discrepancies in coding were discussed between coders until consensus was reached. However, in order to ascertain that training of the coders had been successful, all coders coded the same set of 12 studies and interrater reliabilities were calculated among them. We obtained median interrater reliability of $R = .81$ (mean $R = .69$), which satisfies the usual benchmark for adequate reliability. Moreover, interrater reliability was also calculated between regular coders and the blind coder. Here we obtained slightly lower levels of agreement (median $R = .68$), which was to be expected given that the blind coder coded only conceptual variables that involved a relatively high level of subjectivity. Further details on interrater reliability are available in Supplement 2.

Establishing independent samples. The *summary statistic method* used by the previous major meta-analysis on the predictive validity of the IAT (Greenwald et al., 2009) can result in artificial deflation of variability in effect sizes and moderator variables as well as a loss of information. Therefore, in the current work, we established *independent samples* of interest and explicitly modeled statistical dependencies among effect sizes extracted from the same study (see also Oswald et al., 2013). To determine independent samples, three criteria were used: (*1*) manipulations that may influence the implicit–criterion relationship, such as experimental vs. control condition; (*2*) authors' own hypotheses about the effects of participants' group membership (e.g., male or female, White or Black) on implicit–criterion correlations; and (*3*) participants' group membership (stigmatized vs. nonstigmatized) in the absence of a specific hypothesis.[3] Use of the two latter criteria was driven by the consideration that measures of cognition and behavior may be differentially associated among members of stigmatized vs. nonstigmatized groups. For

[3] This criterion was not used if study authors expressly deemed sample sizes to be too small for splitting the sample.

instance, a measure of identification with science may be correlated with test scores among women but not among men.

Effect sizes. The main effect sizes of interest included the zero-order correlation between implicit and criterion measures (implicit–criterion correlation; ICC), the zero-order correlation between explicit and criterion measures (explicit–criterion correlation; ECC), and the zero-order correlation between implicit and explicit measures (implicit–explicit correlation; IEC). These correlations were retrieved from studies and recoded such that positive values reflected the theoretically expected relationship among implicit, explicit, and criterion measures. That is, ICCs and ECCs were coded as positive if more bias on the implicit or explicit measure was associated with higher levels of discrimination, whereas IECs were coded as positive if higher levels of implicit bias were associated with higher levels of explicit bias.

To be eligible, *implicit measures* had to (*a*) focus on social groups and (*b*) measure attitudes (i.e., associations of social groups with valence such as *male*/*female–good*/*bad*), stereotypes (i.e., associations of social groups with semantic attributes such as *male*/*female–math*/*arts*), or identity (i.e., associations of the self with semantic attributes such as *me*/*not me–math*/*arts*). *Explicit measures* were included if they were parallel to the implicit measure, i.e., if they assessed the same construct. For example, if a study included both a Black/White–good/bad attitude and a Black/White–athletic/intelligent stereotype IAT, the Modern Racism Scale (McConahay, 1986) would be considered an explicit measure only in relation the attitude IAT but not in relation to the stereotype IAT. Finally, all *criterion measures* measured some form of intergroup behavior, in most cases discriminatory behavior.[4] Some studies reported multiple implicit, ex-

[4] For a list of potential criterion measures not included in the present meta-analysis, see Supplement 2.

plicit, or criterion measures; if this was the case, each measure was recorded separately in the dataset.

Moderator variables. A large number of potential moderator variables were coded, including basic study characteristics such as target group, type of criterion behavior, and study setting as well as all conceptual and methodological variables included in previous meta-analyses investigating the relationship between the IAT and behavior (Greenwald et al., 2009; Oswald et al., 2013). Conceptual moderators included the ones derived from dual-process theories such as social sensitivity, controllability, and conscious awareness, as well as ones derived from the study of explicit cognition such as correspondence and target construct (attitude vs. stereotype vs. identity). Methodological moderators included study focus, type of implicit measure (IAT vs. IAT variants), criterion scoring method (relative vs. absolute), and attribute polarity (ranging from low to high). An exhaustive list of moderator variables, along with their definitions and notes on their coding, can be found in Supplement 5.

Analytic strategy

Instead of aggregating dependent effect sizes, statistical analyses explicitly modeled dependencies among effect sizes extracted from the same study. Unless noted, all statistical models were fit using a correlated effects model (Hedges et al., 2010) implemented in the *robumeta* package (Fisher & Tipton, 2015) in the R statistical environment (see also Oswald et al., 2013). All estimates were robust to the assumed within-study correlation between effect sizes. To avoid confusion, regression intercepts are marked b_0 and regression slopes are marked b'. For statistical procedures assuming independent observations, we sampled a single effect size from each study across a large number of iterations and report the median of the resulting distribution of interest.

Results and Discussion[5]

Conceptual and statistical heterogeneity in ICCs

Establishing the degree of heterogeneity in implicit–criterion correlations is of interest for both statistical and conceptual reasons. From a statistical perspective, if effect sizes are found to be heterogeneous, fixed-effects models that assume a single underlying population effect size are inappropriate. Rather, the analytic strategy of choice should be a random-effects model, which allows for meaningful variation in effect sizes beyond sampling variability. Conceptually, high degrees of heterogeneity indicate that instead of asking *whether* implicit measures of intergroup cognition are related to measures of intergroup behavior, it may be more appropriate to ask *under what conditions* the two are more or less highly correlated.

The focal concepts under study in the present work (i.e., *implicit cognition* and *intergroup behavior*) are both quite diffuse. Although all studies included in the meta-analytic database share some methodological features by virtue of using an IAT or related measure, the IATs themselves differ from each other in numerous ways, including target groups and attributes used. As shown in Figure 1, many studies addressed issues of race and ethnicity (k_{ind} = 139)[6] or gender (k_{ind} = 43), with smaller numbers of several other target groups such as sexual orientation and clinically stigmatized groups (total k_{ind} = 75). Moreover, the studies used widely divergent criterion measures such as the expression of policy preferences, resource allocation, academic performance, subtle nonverbal behaviors, performance on interference tasks like the Stroop task,

---

[5] Beyond the main analyses reported in the paper, numerous additional analyses are available in the supplementary materials, which in this case are extensive. Such analyses, along with all materials and the meta-analytic database are available for download from the Open Science Framework (https://osf.io/47xw8/). We hope that other researchers will use these data and materials to reach their own informed conclusions about the relationship between the IAT and intergroup behavior, perhaps using analytic strategies entirely different from our own.

[6] Because the effect sizes included in the meta-analytic database are not statistically independent if they were extracted from the same study, we focus on the number of independent effect sizes, which we denote k_{ind}.

and criminal sentencing decisions. The most frequent criterion categories included person perception (k_{ind} = 53), social affiliation (k_{ind} = 47), and ideology (k_{ind} = 37).

Given the variability in social groups included, types of IATs used, samples studied and, perhaps most strikingly, the variability in criterion variables, effect sizes were expected to show a high degree of statistical heterogeneity. A *prediction interval*, which is a measure of the expected range of effect sizes in a given domain (Borenstein, Higgins, Hedges, & Rothstein, 2017), was calculated. Homogeneous effects result in narrow prediction intervals, whereas heterogeneous effects result in wide prediction intervals.[7] We obtained a 90-percent prediction interval of r_{min} = -.14 to r_{max} = .32, indicting that implicit–criterion associations in the domain of intergroup discrimination should be expected to range from small negative to medium-sized positive relationships. Statistically, the high degree of heterogeneity suggests that any single point estimate of the implicit–criterion relationship would be misleading. Conceptually, it suggests that debates about whether implicit cognition and behavior are related to each other are unlikely to offer any meaningful conclusions. Rather, the focus should be on revealing the conditions under which implicit–criterion correlations are relatively low or high; therefore, we devote most of the remainder of this section to this issue.

Incremental predictive validity

Prior work conducted by Greenwald et al. (2009) used partial correlations to demonstrate that implicit and explicit measures have incremental predictive validity over and above each other. However, as pointed out by Greenwald and colleagues, this approach can be problematic because it ignores measurement error in both variables (for an in-depth explanation and examples

[7] Importantly, unlike a confidence interval, this measure relies on the standard deviation and not the standard error and therefore does not depend on the number of effect sizes included in a meta-analysis. As such, a wide prediction interval does not indicate an imprecise estimate of the mean effect size.

see Westfall & Yarkoni, 2016). Therefore, in the present work we use the structural equation modeling (SEM) approach recommended by Westfall and Yarkoni (2016), which takes into account the internal consistency of both implicit and explicit measures. Implicit measures were found to be associated with criterion measures over and above explicit measures, $\beta = .14$ [.09; .19], $t(145) = 5.41$, $p < .001$. Explicit measures had smaller, but comparable, incremental predictive validity, $\beta = .11$ [.07; .16], $t(145) = 4.65$, $p < .001$. Taken together, these analyses replicate with a stronger method a result also reported by Greenwald et al. (2009): In the intergroup domain, implicit and explicit measures of social cognition each show unique associations of similar magnitude with behavior.[8] The fact that implicit and explicit measures produced similar effect sizes is quite remarkable: Unlike implicit measures and criterion measures, explicit measures and criterion measures were often procedurally similar to each other, which should enhance the observed correlation between the two due to shared method variance (Campbell & Fiske, 1959).

Methodological shortcomings of the reviewed studies

Treatment of measurement error. The latter finding regarding the use of SEM to establish incremental predictive validity highlights the limitations of the analytic strategies used by most primary studies. Given that implicit and explicit measures are correlated with each other and both contain measurement error, most currently used approaches cannot conclusively esti-

[8] Nevertheless, the present findings should be seen as tentative. Because of the low rates of reporting the internal consistency of implicit and explicit measures (20 and 33 percent, respectively), an imputation method had to be used. Specifically, to be able to fit the appropriate structural equation models, we first found the best-fitting beta distribution for each of the two reliability variables. For the reliability of implicit measures, we obtained Beta(4.97, 2.05) and for the reliability of explicit measures, we obtained Beta(15.99, 3.70). We then imputed missing reliability values by drawing random observations from the appropriate beta distribution. To arrive at a stable estimate, this procedure was repeated 1,000 times and the values reported here are the medians of the resulting standardized path coefficient distributions. For $\beta_{implicit}$, mean = .16, $SD = .11$; for $\beta_{explicit}$, mean = .12, $SD = .10$. The relatively high standard deviations indicate relatively high levels of uncertainty around these estimates due to the low reporting rate of the reliability of implicit and explicit measures. Because of this uncertainty, all moderator analyses reported below used zero-order correlations as the dependent variable. To be able to use this method as intended, it is strongly recommended that future research calculate and report measures of internal consistency for both implicit and explicit measures.

mate the magnitude of the implicit–criterion relationship. Zero-order implicit–criterion correlations with (k = 16) or without parallel explicit measures (k = 50), as well as partial correlations (k = 3) and multiple regressions (k = 30) should be superseded by SEM including both implicit and explicit measures as the method of choice in primary research.

Publication bias. The suppression of nonsignificant results can lead to inflated estimates of the meta-analytic effect size and mistaken statistical inferences about the role of moderators (Fanelli & Ioannidis, 2013; Ioannidis, 2008; Rosenthal, 1979). Given that a relatively large number of unpublished effect sizes were included in the meta-analytic database, we were able to probe whether the magnitude of ICCs differed depending on the publication status of the study. No evidence of publication bias was obtained (see Table 1); in fact, effect sizes extracted from unpublished studies were descriptively larger than effect sizes extracted from published work. Second, we fit the three-parameter model of Vevea and Hedges (1995) to all published effect sizes, using a one-sided cutoff at $p < .05$, i.e., assuming a model under which authors selectively report only positive correlations.[9] Addition of this parameter did not lead to any improvement in model fit, $\chi^2(1) = 1.76$, $p = .184$, also providing no evidence of publication bias.[10] Taken together, these results should increase confidence in the validity of the statistical tests reported here and suggest, that unlike other fields of psychology, the study of implicit–criterion relationships is unlikely to be plagued by a widespread file drawer problem.

---

[9] The model was fit using only published effect sizes. Because the method assumes independent observations, we recalculated the likelihood ratio test 1,000 times, each time randomly taking one independent effect size from each study. The χ^2 statistic reported above is the statistic accompanying the median of the resulting distribution of corrected effect size estimates.

[10] A model including five cutoff parameters (ps = .025, .05, .20, .50, and 1, respectively), assuming a more complex and graded selection process, yielded a similar result, $\chi^2(4) = 4.83$, $p = .304$. In Supplement 6, we report additional tests of publication bias, which all converge on the conclusion that the meta-analytic database was unlikely to have been affected by any major suppression of nonsignificant results.

Statistical power. The power of inferential tests has far-reaching consequences for the validity of statistical inferences (Cohen, 1962; Fraley & Vazire, 2014). Therefore, establishing the power of the studies on implicit–criterion correlations is paramount to diagnosing the overall methodological soundness of this literature. The vast majority of the studies included in the present meta-analysis were underpowered: At 40, the median sample size was surprisingly, perhaps shockingly, low. This sample size is miniscule for probing individual differences and too small to reliably (i.e., with a probability of at least .80) detect any effect below the effect size of $r = .43$ (Cohen, 1992). Moreover, a sample size of 40 provides only .40 power to detect the mean effect size reported by Greenwald et al. (2009) and .14 power for the mean effect size reported by Oswald et al. (2013).[11] Even though *post-hoc* power tends to overestimate the power of studies for small effect sizes and small sample sizes (Yuan & Maxwell, 2005), median *post-hoc* power of the included studies was found to be only .15 and mean *post-hoc* power was .27.[12]

These low levels of statistical power are worrisome when it comes to the interpretability and inferential value of the vast majority of individual studies conducted on implicit–criterion relationships. We can go so far as to say that many of the studies included in this meta-analysis should never have been undertaken given the potential for incorrect inferences about the population effect size. Low statistical power of individual studies also provides additional justification for this meta-analysis: Due to their ability to pool data from participants across multiple investi-

[11] Given the high degree of heterogeneity among effect sizes, these power estimates are not necessarily informative. In the Conclusions and Recommendations section we describe an online tool that can be used to calculate the expected effect size and the sample size required to achieve adequate power given a certain configuration of study characteristics.

[12] Similar results were obtained for the subset of studies whose main focus was the relationship between implicit social cognition and intergroup behavior. Median sample size in this subset of studies was 45. The smallest correlation for whose detection this sample size provides adequate power is $r = .41$. Median *post-hoc* power was .16 and mean *post-hoc* power was .29.

gations, meta-analyses have the potential to derive valid conclusions about the population effect size and its moderators even when individual studies are underpowered (e.g., Card, 2016).

Basic study characteristics: Target group, type of behavior, and study setting

Target group. Regarding the target group variable, two results seem noteworthy (see Figure 1). First, implicit attitudes were significantly associated with behavior across all target categories, with the exception of one category labeled "other intergroup," which was highly diverse and contained a relatively small number of effect sizes (k_{ind} = 19). Importantly, this result indicates that implicit–criterion correlations (ICCs) were fairly homogeneous across target group categories. On the other hand, explicit–criterion correaltions (ECCs) were found to be more variable by target group than ICCs. For the former, effect sizes ranged from r = .10 (ethnicity) to r = .32 (sexuality), whereas for the latter they ranged from r = .08 (other clinical) to r = .11 (sexuality).

Type of criterion behavior. Implicit measures were found to be associated with behavior across the vast majority criterion behavior categories. The four exceptions out of 18 criterion categories included cognitive measures, neurological and physiological measures, prosocial behavior, and resource allocation. However, three out of these four categories contained fewer than 10 studies, which leads to relatively uncertain estimates and, possibly, lack of statistical significance. Paralleling the effects of target group, the range of variation in ECCs across criterion categories was considerably larger than the range of variation in ICCs. The estimated category means for ECCs ranged from r = -.06 for neurological and physiological measures to r = .45 for group perception, compared to a much narrower range of r_{min} = -.05 (resource allocation) to r_{max} = .16 (cognitive measures) for ICCs. ECCs were not significantly different from zero for nine out of 18 categories, as opposed to four for ICCs. However, this difference may, at least in part, be

due to the fact that not all studies contained parallel explicit measures and thus the degrees of freedom for models with ECC as the dependent variable were smaller. With the caveats mentioned above, these findings suggest that whereas implicit measures tend to produce small but consistent effects across target group and criterion categories, explicit measures are strongly related to certain kinds of intergroup behavior but unrelated to others.

Study setting. Given considerable levels of interest in using the IAT for prediction in real-world settings, we also probed whether effect sizes significantly differed as a function of study setting, i.e., across (*a*) lab studies conducted mostly with undergraduate participants, (*b*) real-world studies conducted in field settings, and (*c*) studies conducted over the Internet with online participants. Even though the latter two categories contained fairly large numbers of effect sizes (k_{ind} = 23 and 39, respectively), ICCs were significantly different from zero and highly similar in size across all study locations (see Table 1). More work will be necessary to fully explore the implicit–criterion relationship in real-world contexts. However, given a sizeable number of effect sizes in each category, it is warranted to conclude that the relationship between implicit measures and intergroup behavior is sufficiently robust to emerge not only in the lab but also under the less controlled conditions of online and field studies.

Conceptual variables

Effects on ECCs. The higher the controllability of the behavior, the higher participants' conscious awareness that discrimination was being measured, and the higher the correspondence between cognition and behavior, the larger explicit–criterion correlations became (see Table 2). At the .10 quantile of the controllability variable, the predicted ECC was .01, whereas at the .90 quantile, it increased to .16. Similarly, for conscious awareness, predicted ECC increased from .07 to .17, and for correspondence, from .09 to .17. In addition, in line with the previous meta-

analysis by Talaska et al. (2008), we found that explicit attitudes (mean $r = .15$) were more highly correlated with behaviors than explicit stereotypes (mean $r = .08$). Overall, these results provide evidence for the theories of explicit–behavior relationships reviewed above and replicate previous findings.

Unexpectedly, and contrary to the results obtained by Greenwald et al. (2009), social sensitivity did not emerge as a moderator of ECCs. This finding may be interpreted in a number of ways. First, the range of the social sensitivity variable was more restricted here given that the present meta-analysis investigated only intergroup discrimination, whereas Greenwald and colleagues also included areas such as consumer behavior and close relationships. Second, the social sensitivity variable itself may have been subject to impression management pressures. That is, coders may have been reluctant to admit that it is socially acceptable to express negative attitudes toward or negative stereotypes of certain groups.[13] Finally, social sensitivity and ECCs may have been unrelated because more socially sensitive explicit measures may, on average, have been used to predict more socially sensitive behaviors. To the extent that this is the case, both explicit and criterion measures may have been affected by the same impression management pressures and therefore ECCs may have remained high even at high levels of social sensitivity.

Effects on ICCs. The results obtained with ICCs as the response variable stand in stark contrast to the analyses involving ECCs, given that ICCs were not consistently predicted by any

[13] In line with this explanation, the means of both social sensitivity variables (the one for the explicit and the one for the implicit measure, respectively) were well above the midpoint of 7-point the scale (5.37 and 5.23, respectively). In spite of the relatively high means, the coding of the social sensitivity variables did reflect commonsense notions about different levels of social sensitivity being associated with different target groups. For instance, the highest social sensitivity values were assigned to race (means of 6.19 and 6.55), with lower values for gender (means of 4.28 and 4.02), and even lower values for age (means of 2.57 and 3.20). However, groups with higher levels of social sensitivity (race and ethnicity) were overrepresented among the included studies, which explains the relatively high overall means.

of the conceptual variables included (see Table 2).[14] That is, univariate meta-regressions showed that implicit measures were equally associated with measures of intergroup behavior irrespective of social sensitivity, controllability, conscious awareness, or target concept. In fact, contrary to the widespread notion that implicit measures are not associated with highly controllable behaviors, the present meta-analysis found a sizable number of large ICCs for such behaviors, including self-reported enrollment intentions in mathematics classes (Steffens, Jelenec, & Noack, 2010), self-reported career aspirations (Asgari, Dasgupta, & Stout, 2012), and voting behavior (Bernstein, Young, & Claypool, 2010). As such, these results suggest that the contents of implicit social cognition may be more accessible to conscious introspection than usually assumed (De Houwer, 2006). Alternatively, less consciously activated representations might influence any type of behavior irrespective of its controllability or amenability to conscious awareness.

The only conceptual variable that emerged as a significant moderator of ICCs is *correspondence*, suggesting that implicit measures may be more closely associated with behaviors with which they share causal antecedents. As correspondence increased from the .10 to the .90 quantile, predicted ICC increased from .05 to .14. However, this result should be interpreted with caution given that, unlike all other results discussed above, the models involving regular and

---

[14] Using a simulation approach to establish the power of the univariate regression analyses reported here, we found that the models had .80 power to detect an effect size of $r = .13$, .90 power to detect $r = .14$, and .99 power to detect $r = .18$. For the power analysis, we simulated variables that reflected the average within-study dependency among the moderators in our dataset (mean R^2 based on weighted regressions = .90) and the average within-study dependency among ICCs ($R^2 = .42$). Then, in each iteration of the simulation, we generated a normal variable representing a moderator and another normal variable representing the ICC. Across simulations, we varied the correlation between the moderator and the ICC (r ranging from 0 to .55), with 1,000 iterations for each effect size. Then we fit a univariate meta-regression to each pair of stimulated variables using the sample sizes and regression weights from the observed data and took the percentage of significant effects within each effect size as the power of the meta-regression to detect that effect size.
These obtained power values seem adequate given that even $r = .18$ is usually considered to be a small effect (Cohen, 1992). At $r = .07$, the observed median correlation between the significant moderators of interest and ICCs was lower than the threshold for achieving .80 power. However, power was fair if below desired levels for most moderators of interest, including study focus, correspondence, and attribute polarity (ps = .61).

blind-coded variables suggested different conclusions. Whereas the former showed significant moderation, the latter did not. Therefore, the possibility that knowledge of the study outcomes may have influenced coding of the correspondence variable among regular coders cannot be ruled out at this time. Moreover, further theoretical and empirical work may be necessary to understand how correspondence (a construct developed to describe the relationship between self-reported attitudes and behavior) applies to implicit social cognition.

IEC–ICC and IEC–ECC relationships. We obtained a positive relationship between IECs and ICCs, $b' = .23$ [.08; .37], $t(149) = 3.05$, $p = .003$. That is, implicit measures were found to be more closely associated with criterion behaviors when implicit and explicit measures were highly correlated. The same was not true for ECCs; that is, the predictive power of explicit measures did not depend on the strength of the implicit–explicit relationship, $b' = .04$ [-.14; .21], $t(145) = 0.41$, $p = .685$. This finding seems to support theorizing by Greenwald and Banaji (2017) and puts on the table a new hypothesis deserving of more thorough empirical testing: the possibility that implicit cognition influences explicit cognition that, in turn, drives behavior.

To summarize, conceptual moderators, such as controllability, conscious awareness, and target concept, significantly affected explicit–criterion relationships. By contrast, correlations between implicit and criterion measures remained robust irrespective of conceptual moderators. This suggests that existing dual-process theories may need to be revised to account for the conditions under which implicit measures and intergroup behavior are more or less strongly related to each other. Moreover, also contrary to dominant dual-process theories, but in line with recent theorizing by Greenwald and Banaji (2017), we found that higher correlations between implicit and explicit measures predicted higher implicit–criterion correlations (but not higher explicit–criterion correlations).

Methodological moderators

The conceptual variables derived from existing dual-process theories and theories of explicit social cognition failed to reliably account for any of the variance in ICCs. However, design-related moderators, such as study focus, criterion scoring method, type of implicit measure, and attribute polarity, produced significant effects on the implicit–criterion relationship (for mean ICC estimates see Table 1). These factors are of importance because they are easily controlled by researchers, and as an increasing number of methodological criteria are fulfilled, the sample size required to achieve adequate power becomes considerably smaller (see below). This consideration is paramount because, given current subpar levels of statistical power in the majority of studies, ensuring appropriate sample sizes is a crucial area of potential improvement for future work.

Study focus. Studies whose abstracts highlighted the relationship between implicit measures and behavior as their main focus produced higher effect sizes than those in which this relationship was secondary or incidental.[15] However, ICCs were found to be positive and significantly different from zero in all categories of study focus, except for the one containing the smallest number of effect sizes (k_{ind} = 16). This finding is open to two interpretations. On the one hand, when investigating ICCs is the main focus of a study, authors may be more careful in creating ideal conditions to reveal an implicit–criterion relationship should one exist in the population. On the other hand, the effect of study focus on ICC magnitudes may be, at least in part, reflective of HARKing (Kerr, 1998). That is, once a significant association with behavior was

[15] The meta-regressions reported in Table 2 are intercept-only models that show the mean level of ICC for each level of categorical moderators. These models cannot form the basis of any inferences about differences between these categories. When we mention significant moderation in the text, we do so on the basis of models that explicitly include such comparisons (see Supplement 12).

found, authors may have been more inclined to highlight this finding as a focal result. These two possibilities may be disambiguated by preregistration (Wagenmakers et al., 2012).

Type of implicit measure. Even though ICCs were significantly different from zero for all measures included, the standard IAT (Greenwald et al., 1998) emerged as superior to a combined category of IAT variants, including the SC-IAT (Karpinski & Steinman, 2006), the brief IAT (Sriram & Greenwald, 2009), the GNAT (Nosek & Banaji, 2001), the personalized IAT (M. A. Olson & Fazio, 2004), and the paper-and-pencil IAT (Sekaquaptewa, Vargas, & von Hippel, 2010). The IRAP, at least descriptively, outperformed the IAT (Barnes-Holmes et al., 2006), but the small number of studies involving this measure (k_{ind} = 4 compared to k_{ind} = 217 for the standard IAT) does not allow a conclusion of its superiority to be ascertained with confidence at this time.

Attribute polarity. IATs differ from each other along the dimension of *attribute polarity*, i.e., the extent to which the two attribute concepts are polar opposites of each other. Each implicit measure was assigned to one of three attribute polarity values: 3 (high polarity such as *fat* vs. *thin*), 2 (medium polarity such as *math* vs. *language*), and 1 (low polarity such as *sad* vs. *angry*). We found a positive relationship between attribute polarity and ICCs: IATs with high-polarity attributes were more highly correlated with criterion behaviors, presumably because such IATs tap into a more cohesive network of mental representations.

Criterion scoring method. The scoring method used for criterion measures also moderated the magnitude of ICCs (see also Oswald et al., 2013): Relative ratings and difference scores (such as donations to a Black compared to a White student group), were found to be associated with larger effect sizes than absolute ratings, especially ones involving nonstigmatized targets (such as donations to a White student group). This result suggests that implicit measures may be

primarily indicative of differences in behavioral tendencies toward members of stigmatized vs. nonstigmatized groups, a finding with possible theoretical implications (Carlsson & Agerström, 2016).

Different methodological choices produce highly divergent ICCs

To demonstrate the combined effects of the significant moderators of implicit–criterion correlations discussed above, we recalculated the meta-analytic effect size using intercept-only meta-regressions involving different combinations of these moderators. Out of the 2,240 ICCs included in this meta-analysis, there were only 24 effect sizes from 13 studies that (*a*) had the relationship between implicit cognition and behavior as their primary focus, (*b*) used relative or difference score measures of behavior, (*c*) used an IAT or IRAP, (*d*) used attributes that were polar opposites of each other, and (*e*) used highly correspondent implicit and criterion measures. For this subgroup of effect sizes, we found a mean ICC of $r = .37$ [.18; .54], $t(12) = 4.08$, $p = .002$, which can be reliably detected using a relatively small sample of 54 participants.

As criteria were systematically relaxed to include a wider range of studies, the estimate of the effect size decreased and so did the power of the average study to give rise to meaningful inferences about the underlying population effect. As a first step, studies using implicit measures with attributes that are not polar opposites of each other (e.g., *pleasant–threatening* or *sad–angry*) and using criterion measures scored in an absolute manner, i.e., without comparison to another group (e.g., money donated to a Black student organization) were included in the estimate of the effect size. As a result, mean ICC dropped to $r = .23$ [.14; .31], $t(36) = 5.68$, $p < .001$, an effect size whose reliable detection requires a sample size of 145. Further, relaxing the requirements for high correspondence between implicit and criterion measures and using a standard IAT or IRAP resulted in an even lower ICC estimate of $r = .13$ [.10; .16], $t(110) = 8.35$, $p <$

.001 (required $N = 461$). Finally, once all effect sizes from the meta-analytic database were considered regardless of meeting any of the criteria described above, the implicit–criterion correlation became small but remained statistically significant, $r = .10$ [.08; .11], $t(252) = 10.99$, $p < .001$ (required $N = 782$). This result is perhaps the most significant discovery of the present meta-analysis: The stronger the study on the aforementioned methodological variables, the larger the implicit–criterion correlation.

Conclusions and Recommendations

Using data from 217 research reports and a total of over 36,000 participants, a six-fold increase over previous meta-analyses, this project investigated the conceptual and methodological conditions under which Implicit Association Tests measuring attitudes, stereotypes, and identity correlate with criterion measures of intergroup discrimination. Overall, we found (*a*) high levels of heterogeneity in implicit–criterion relationships, (*b*) unique and similarly sized effects of implicit and explicit measures on behavior, (*c*) associations of implicit measures with behavior irrespective of (*i*) basic study characteristics such as target group, type of criterion behavior, and study setting or (*ii*) conceptual moderators such as controllability and conscious awareness of the behavior, (*e*) significant variation in explicit–criterion relationships as a result of both basic study features and conceptual variables such as controllability and conscious awareness, (*f*) higher implicit–criterion correlations when implicit and explicit measures were more highly correlated, and (*g*) significant variation in implicit–criterion relationships as a result of methodological moderators including study focus, type of implicit measure used, measurement of criterion behavior, attribute polarity, and possibly implicit–criterion correspondence.

The difficulty of theorizing about implicit cognition

This meta-analysis found conceptual variables to be reliable predictors of explicit–criterion, but not of implicit–criterion correlations. The absence of theoretical predictors of ICCs, in the presence of several methodological predictors, suggests that theorizing about implicit cognition is relatively unsophisticated at this time. Why might this be? First and foremost, implicit cognition by its nature refers to aspects of human thought that are relatively less accessible to conscious awareness. As such, scientists themselves have a harder time generating good intuitions about the mechanisms of implicit cognition. One can see this in current theorizing that stems from dual-process accounts and posits that indirect measures capture unconscious cognition and should predict automatic behaviors, whereas direct measures capture conscious cognition and should predict controllable behaviors. This thinking may need to be relinquished. One result that emerged from this meta-analysis, replicating previous work by Greenwald et al. (2009), may guide new theorizing: If implicit–criterion correlations are higher the better correlated implicit and explicit measures are, the intriguing suggestion is that implicit cognition makes its way subtly into explicit cognition, through which it influences behavior (see Greenwald & Banaji, 2017).

A second way to think about improving theorizing about implicit–criterion correlations is to search for surprising patterns of data and, provided that they replicate and generalize, use them as a starting point to revise theories. Take, for instance, the received wisdom that the greater the racial or ethnic diversity of a region, the lower the implicit bias toward minorities should be. In fact, as revealed by work conducted by Rae et al. (2015), the opposite is the case: U.S. states with higher proportions of African American residents exhibit *higher*, rather than lower, levels of implicit race bias. Empirical discoveries like this have the potential to inspire the development of new theory in a bottom-up, data-driven, way. They force a reckoning with the evidence to

change theories of implicit–criterion relationships rather than coding them as failures to confirm existing theory (as we did in this meta-analysis).

Recommendations for study design

This meta-analysis revealed that several aspects of research on the cognition–behavior relationship in intergroup contexts are in need of basic methodological improvement. We formulate specific recommendations intended to benefit (*a*) scientists in designing and implementing research and (*b*) editors and funding agencies in evaluating research. Besides discussing the core issues here, we make recommendations available as a succinct 8-point checklist in Supplement 9.

Calculate statistical power given the expected effect size. Although too obvious a point to mention, inadequate statistical power can result in inflated false positive rates, overestimated effect sizes, and low rates of reproducibility (Button et al., 2013). Future research should adhere at least to the reasonable standard of conducting formal power calculations to determine sample size. However, given significant heterogeneity in implicit–criterion correlations, power calculations are non-trivial. Therefore, we developed and present an easy-to-use online power calculator available at http://www.benedekkurdi.com/#iat to estimate the sample size necessary to achieve appropriate levels of statistical power for different, user-specified, configurations of study-specific and methodological variables. For instance, the power calculator allows one to determine the sample size necessary to achieve adequate power in a study using a standard IAT and a relative measure of criterion behavior in the domain of academic performance with sexual orientation as the target category ($N = 162$).

Ensure reliability of implicit measures. Most studies included in this meta-analysis used zero-order correlations or regression models to make claims about implicit–criterion relationships. However, such claims may be misguided because implicit and explicit measures are

significantly correlated with each other and neither is perfectly reliable. Therefore, we strongly recommend that future studies (*a*) include parallel explicit measures along with implicit measures, (*b*) report the internal consistency of all independent variables involved in making claims about associations between measures of cognition and behavior, and (*c*) use adequate statistical methods to adjust for measurement error, such as the structural equation modeling approach recommended by Westfall and Yarkoni (2016). We offer an online tool at http://www.benedekkurdi.com/#iat, which estimates internal consistency based on trial-level IAT data provided by the user.[16]

Focus equal attention on the reliability and validity of *criterion* measures (as on measures of cognition). Decades of research on attitudes and beliefs have sought to establish the validity and reliability of explicit and implicit measures (Banaji, 2001; Thurstone, 1928). If significant progress is to be made on understanding attitude–behavior relationships, the issue of whether behavioral measures successfully capture the intended underlying constructs must receive closer attention (Carlsson & Agerström, 2016; Talaska et al., 2008). Any perusal of this literature will easily reveal the *ad-hoc* nature of criterion measure choices whose validity and reliability has never been established. Indeed, far more effort has been put into establishment of reliability and validity of both explicit and implicit social cognition measures than into the behaviors they are called into service to predict. This must be remedied (for an initial example see Axt, Nguyen, & Nosek, 2018).

Taking stock and future directions

Among the positive recent improvements in exploring implicit–criterion relationships is the presence of a sizeable number of new studies conducted in real-world ($k = 23$) and online

[16] The details of the statistical approach used by the online tool are described in Supplement 10.

contexts (k = 29). Interestingly, this meta-analysis found no difference in average implicit–criterion correlations as a function of study settings. We believe that this is a notable result showing that the implicit–criterion relationship is sufficiently robust so as to be revealed even under the relatively less controlled conditions of the natural world. Moreover, a recent theoretical contribution by Payne, Vuletich, and Lundberg (2017) has directed attention to a growing set of studies in which implicit attitudes, as averaged at the level of a geographic area, are significantly correlated with aggregate outcomes at the level of the same geographic area. For instance, the higher the country-wide implicit gender stereotypes, the less well girls perform relative to boys on an 8th-grade standardized test and the higher the implicit race bias in a geographic region in the U.S., the greater the lethal use of force against African Americans by police. Such studies tend to produce larger effects than studies that correlate implicit measures with behavior at the level of individuals. This suggests that correlations between implicit measures and behavior may, at least in part, arise due to situational factors rather than stable individual differences. A future meta-analysis of aggregate-level prediction studies, most of which were not eligible for inclusion in the present project because they had been published after our cutoff date, will have to ascertain the robustness of this finding. Moreover, disentangling the additive, or perhaps interactive, effects of persons and situations on implicit cognition using repeated measurement designs is another major outstanding challenge to be addressed in future work (Kurdi & Banaji, 2017).

So far, we have provided several reasons for the size of effects obtained in this literature, including insufficient power and shortcomings in analytic strategies. In addition, we pointed out that when studies adhere to particular methodological criteria, ICCs rise substantially, not to mention large effect sizes when implicit measures and measures of behavior are aggregated across a region (Payne et al., 2017). Similarly, effect sizes tend to be relatively large in studies

on physician–patient interactions, in which (like in the studies reviewed by Payne and colleagues), behaviors are aggregated across individuals, thus reducing measurement error (Penner et al., 2010). In conclusion, new types of data collection and analysis, now possible due to open access to large amounts of data, have revealed larger effect sizes than small-scale studies conducted mostly in laboratory contexts. This suggests that as individual-level studies meet increasingly strict methodological criteria and approximate the aggregate-level studies described by Payne and colleagues, we may see an increase in effect sizes in these studies as well.

However, even if effect sizes are relatively small in particular studies, the correlation may still translate into societally meaningful impact given repeated interactions (Greenwald, Banaji, & Nosek, 2015). To illustrate, there were 5.7 million job openings in the United States in March 2017 (Bureau of Labor Statistics, 2017). If one assumes that an equally qualified man and woman applied for each job, an $r = .10$ relationship between implicit gender bias and hiring decisions would result in the hiring of 3.36 million male candidates compared to 2.34 female candidates, a difference of considerable magnitude and practical significance.[17] In this regard, we fully agree with Yarkoni and Westfall (2017) who argue that, *ceteris paribus*, small effects obtained from large samples should always be favored over large effects obtained from small samples.

The data from this meta-analysis leave little doubt about the conclusion that attitudes, stereotypes, and identity, measured both using self-report and less controllable responses such as the Implicit Association Test, are systematically related to behavior in the intergroup domain. But why might this relationship arise? On the one hand, implicit attitudes vary as a function of relatively stable features of individuals and the environment such as group membership, political

[17] A point-biserial correlation of $r = .10$ corresponds to an odds ratio of 1.44, which, in turn, translates into a probability of $p = .59$ of the male candidate being hired and a corresponding probability of $1 - p = .41$ of the female candidate being hired. 5.7 million multiplied by p gives 3.36 million.

orientation, and geography (Charlesworth & Banaji, 2018; Nosek et al., 2007). On the other hand, implicit attitudes are also sensitive to relatively transient contextual features, such as shifts in the immediate environment in which the test is taken (Lai, Hoffman, & Nosek, 2013). As with any measure of attitudes or other individual differences, the IAT score is not a static measure of disposition: It is an adaptive response produced in a particular situation by an organism with a particular biology, personality, and cultural history (Mischel, 1968). Given such malleability, we have always advised against using a single intergroup IAT as a device for the selection of people, such as whether to hire someone for a job or admit them to a club. The measure is of value in two contexts: research and education. The Project Implicit educational website (http://implicit.harvard.edu/) has served primarily as a platform for education about the hidden aspects of mental processes. Moreover, the substantial use of the IAT in research in a wide range of fields including clinical, organizational, educational, medical, business, and legal contexts demonstrates its viability. However, it is our hope that a better method will replace the IAT, and we urge such development.

Given the large number of studies available in the intergroup domain alone, this meta-analysis was selected to produce a quantitative summary of this field. However, a wealth of work has been generated on the relationship between implicit measures and behavior in other domains such as the diagnosis of clinical conditions, consumer preferences, and political behavior. Our hope is that in the future new meta-analyses will be conducted in these and other areas, and their results will inform and constrain the findings obtained in the present project. Specifically, it would be important to investigate whether the theoretical findings derived from this review of studies in the area of intergroup relations are applicable to implicit social cognition in general. For now, we look forward to new empirical evidence being created in accordance with the meth-

odological recommendations prescribed here. Such new studies will no doubt go a long way toward furthering our understanding of the relationship between implicit cognition and behavior in the intergroup domain.

References

Studies included both in the text of the article and the meta-analytic database are marked with an asterisk (*). An exhaustive list of all articles included in the meta-analysis is available in Supplement 11.

* Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805. http://doi.org/10.1037/a0021594

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. http://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*(4), 652–661. http://doi.org/10.1037/0022-3514.91.4.652

* Asgari, S., Dasgupta, N., & Stout, J. G. (2012). When do counterstereotypic ingroup members inspire versus deflate? The effect of successful professional women on young women's leadership self-concept. *Personality and Social Psychology Bulletin*, *38*(3), 370–383. http://doi.org/10.1177/0146167211431968

Axt, J. R., Nguyen, H., & Nosek, B. A. (2018). The Judgment Bias Task: A flexible method for assessing individual differences in social judgment biases. *Journal of Experimental Social Psychology*. Advance online publication. http://doi.org/10.1016/j.jesp.2018.02.011

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger III, J. S. Nairne, & I. Neath (Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder* (pp. 117–149). Washington, D.C.: American Psychological Association.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. http://doi.org/10.3758/s13428-013-0410-6

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*, 169–177.

* Bernstein, M. J., Young, S. G., & Claypool, H. M. (2010). Is Obama's win a gain for Blacks? Changes in implicit racial prejudice following the 2008 election. *Social Psychology*, *41*(3), 147–151. http://doi.org/10.1027/1864-9335/a000021

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. http://doi.org/10.1002/jrsm.1230

Bureau of Labor Statistics. (2017, March 9). Job Openings and Labor Turnover Summary. Retrieved May 18, 2017, from https://www.bls.gov/news.release/jolts.nr0.htm

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. http://doi.org/10.1038/nrn3475

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. http://doi.org/10.1037/h0046016

Card, N. A. (2016). *Applied meta-analysis for social science research*. New York, NY: Guilford.

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology*, *57*(4) 278–287. http://doi.org/10.1111/sjop.12288

Charlesworth, T. E. S., & Banaji, M. R. (2018). *Patterns of implicit and explicit attitudes I. Long-term change and stability from 2007–2016.* Manuscript submitted for publication.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153. http://doi.org/10.1037/h0045186

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://doi.org/10.1037/0033-2909.112.1.155

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. http://doi.org/10.1016/j.lmot.2005.12.002

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. http://doi.org/10.1037//0022-3514.56.1.5

Dunham, Y., Chen, E. E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science*, *24*(6), 1–27. http://doi.org/10.1177/0956797612463081

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*(37), 15031–15036. http://doi.org/10.1073/pnas.1302997110

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. http://doi.org/10.1146/annurev.psych.54.101601.145225

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. http://doi.org/10.1037/0022-3514.50.2.229

Fisher, Z., & Tipton, E. (2015, March 7). robumeta: An R-package for robust variance estimation in meta-analysis. Retrieved from http://arxiv.org/abs/1503.02220

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10), e109019–12. http://doi.org/10.1371/journal.pone.0109019

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*(1), 23–30. http://doi.org/10.2307/3033657

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/10.1037//0033-295X.102.1.4

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*(9), 861–871. http://doi.org/10.1037/amp0000238

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. http://doi.org/10.1037/pspa0000016

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differ-

ences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. http://doi.org/10.1037//0022-3514.74.6.1464

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*(1), 17–41. http://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*(1), 39–65. http://doi.org/10.1002/jrsm.5

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. http://doi.org/10.1177/0146167205275613

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. http://doi.org/10.1097/EDE.0b013e31818131e7

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16–32. http://doi.org/10.1037/0022-3514.91.1.16

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. http://doi.org/10.1207/s15327957pspr0203_4

* Krieger, N., Waterman, P. D., Kosheleva, A., Chen, J. T., Carney, D. R., Smith, K. W., et al. (2011). Exposing racial discrimination: Implicit & explicit measures–The my body, my story study of 1005 US-born Black & White community health center members. *PLoS ONE*, *6*(11), e27636–25. http://doi.org/10.1371/journal.pone.0027636

Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, *28*(4), 281–287. http://doi.org/10.1080/1047840X.2017.1373555

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*(5), 315–330. http://doi.org/10.1111/spc3.12023

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. http://doi.org/10.1037/a0036260

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces*, *13*(2), 230–237. http://doi.org/10.2307/2570339

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). San Diego, CA: Academic Press.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. http://doi.org/10.1037/h0031564

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. http://doi.org/10.3758/BF03213230

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625–644. http://doi.org/10.1521/soco.19.6.625.20886

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. http://doi.org/10.1177/0146167204271418

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. http://doi.org/10.1080/10463280701489053

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*(5), 653–667. http://doi.org/10.1037//0022-3514.86.5.653

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*(2), 171–192. http://doi.org/10.1037/a0032734

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233–248. http://doi.org/10.1080/1047840X.2017.1335568

* Penner, L. A., Dovidio, J. F., West, T. V., Gaertner, S. L., Albrecht, T. L., Dailey, R. K., & Markova, T. (2010). Aversive racism and medical interactions with Black patients: A field study. *Journal of Experimental Social Psychology*, *46*(2), 436–440. http://doi.org/10.1016/j.jesp.2009.11.004

* Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activa-

tion. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. http://doi.org/10.1162/089892900562552

* Rae, J. R., Newheiser, A.-K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science*, *6*(5), 535–543. http://doi.org/10.1177/1948550614567357

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. http://doi.org/10.1037/0033-2909.86.3.638

* Rüsch, N., Todd, A. R., Bodenhausen, G. V., Olschewski, M., & Corrigan, P. W. (2010). Automatically activated shame reactions and perceived legitimacy of discrimination: A longitudinal study among people with mental illness. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(1), 60–63. http://doi.org/10.1016/j.jbtep.2009.10.002

Sekaquaptewa, D., Vargas, P., & von Hippel, W. (2010). A practical guide to paper and pencil implicit measures of attitudes. In B. Gawronski & B. K. Payne (Eds.), *Handbook of Implicit Social Cognition: Measurement, theory and applications* (pp. 140–155). New York, NY: Guilford Press.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*(4), 283–294. http://doi.org/10.1027/1618-3169.56.4.283

* Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, *102*(4), 947–963. http://doi.org/10.1037/a0019920

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247.

http://doi.org/10.1207/s15327957pspr0803_1

Swann, W. B., Jr., & Bosson, J. K. (2010). Self and identity. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 589–628). Hoboken, NJ: John Wiley & Sons, Inc.

Talaska, C. A., Fiske, S. T., & Chaiken, S. (2008). Legitimating racial discrimination: Emotions, not beliefs, best predict discrimination in a meta-analysis. *Social Justice Research*, *21*(3), 263–296. http://doi.org/10.1007/s11211-008-0071-2

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–27. http://doi.org/10.1086/214483

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. http://doi.org/10.1007/BF02294384

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. http://doi.org/10.1177/1745691612463078

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, *11*(3), e0152719–22. http://doi.org/10.1371/journal.pone.0152719

Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, *25*(4), 41–78. http://doi.org/10.1111/j.1540-4560.1969.tb00619.x

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. http://doi.org/10.1037//0033-295X.107.1.101

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *11*, 174569161769339–23. http://doi.org/10.1177/1745691617693393

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*(2), 141–167. http://doi.org/10.3102/10769986030002141

Table 1. Summary table of univariate meta-regressions predicting implicit–criterion correlations (ICCs) based on methodological moderators. k_{total} = total number of effect sizes included in the model, k_{ind} = number of independent effect sizes included in the model, *b* = unstandardized regression coefficient, CI lower = lower bound of the confidence interval, CI upper = upper bound of the confidence interval, DF = degrees of freedom, *t* = value of the *t* statistic, *p* = *p* value, τ^2 = residual heterogeneity. For categorical predictors (with their levels listed), *b* coefficients represent condition means, whereas for metric predictors, *b* coefficients represent units of change in the dependent variable (ICC) associated with one unit of change in the moderator variable.

Moderator	k_{total}	k_{ind}	*b*	CI lower	CI upper	DF	*t*	*p*	τ^2
Study-level moderators									
Publication status									
Unpublished	177	27	0.146	0.070	0.221	26	3.91	.001	.047
Published	2063	226	0.090	0.074	0.107	225	10.65	.001	.018
Study location									
Lab	1731	202	0.100	0.079	0.121	201	9.34	.001	.020
Real-world	196	23	0.092	0.044	0.139	22	4.01	.001	.023
Online	313	29	0.086	0.047	0.126	28	4.45	.001	.023
Study focus									
Primary w/o moderator	810	111	0.130	0.100	0.161	110	8.35	.001	.028
Primary with moderator	453	48	0.091	0.062	0.120	47	6.25	.001	.016
Secondary w/o moderator	190	17	0.068	0.009	0.126	16	2.45	.026	.010
Incidental	641	63	0.053	0.023	0.083	62	3.54	.001	.013
Secondary with moderator	146	16	0.044	-0.016	0.103	15	1.57	.138	.015
Measure-level moderators									
Type of implicit measure									
IRAP	14	4	0.207	0.012	0.387	3	3.38	.043	.027
IAT	1881	217	0.101	0.082	0.120	216	10.22	.001	.020
IAT variant	345	37	0.055	0.022	0.089	36	3.38	.002	.024
Attribute polarity	2184	250	0.073	0.005	0.141	247	2.11	.036	.021
Criterion scoring method									
Relative rating	101	24	0.151	0.057	0.242	23	3.31	.003	.072
Difference score	294	58	0.133	0.076	0.190	57	4.61	.001	.029
Single rating stigmatized	1212	202	0.085	0.069	0.101	201	10.39	.001	.012
Single rating non-stigmatized	633	93	0.051	0.022	0.081	92	3.43	.001	.019

Table 2. Summary table of univariate meta-regressions predicting implicit–criterion correlations (ICCs) and explicit–criterion correlations (ECCs) on the basis of conceptual moderators. Mean = mean of moderator variable, *SD* = standard deviation of predictor variable, k_{total} = total number of effect sizes included in the model, k_{ind} = number of independent effect sizes included in the model, *b* = unstandardized regression coefficient, CI lower = lower bound of the confidence interval, CI upper = upper bound of the confidence interval, DF = degrees of freedom, *t* = value of the *t* statistic, *p* = *p* value, τ^2 = residual heterogeneity, [B] = blind-coded version of moderator variable. For categorical predictors (with their levels listed), *b* coefficients represent condition means, whereas for metric predictors, *b* coefficients represent units of change in the dependent variable (ICC or ECC) associated with one unit of change in the moderator variable.

Moderator	Mean	*SD*	k_{total}	k_{ind}	*b*	CI lower	CI upper	DF	*t*	*p*	τ^2
Implicit–criterion correlation (ICC)											
Implicit social sensitivity	5.375	1.598	2233	252	0.004	-0.005	0.014	250	0.85	.396	.021
Implicit social sensitivity [B]	4.635	1.738	2240	253	0.003	-0.007	0.013	251	0.54	.590	.021
Controllability	7.742	3.339	2240	253	0.000	-0.008	0.008	251	-0.11	.914	.020
Controllability [B]	6.400	2.825	2240	253	-0.001	-0.010	0.008	251	-0.21	.832	.020
Awareness	6.853	3.253	2240	253	-0.002	-0.007	0.004	251	-0.54	.588	.020
Awareness [B]	5.537	3.066	2240	253	0.002	-0.005	0.008	251	0.50	.620	.020
Correspondence	2.057	0.653	2240	253	0.047	0.020	0.074	251	3.40	.001	.020
Correspondence [B]	1.608	0.660	2240	253	0.016	-0.008	0.039	251	1.30	.195	.020
Concept											
Stereotype	–	–	739	76	0.098	0.073	0.124	75	7.64	.001	.011
Attitude	–	–	1333	189	0.096	0.074	0.117	188	8.60	.001	.025
Identity	–	–	168	22	0.086	0.024	0.147	21	2.89	.009	.017
Explicit–criterion correlation (ECC)											
Explicit social sensitivity	5.231	1.660	1735	158	-0.007	-0.023	0.010	156	-0.80	.427	.056
Explicit social sensitivity [B]	4.724	1.787	1739	160	0.001	-0.014	0.015	158	0.10	.920	.056
Controllability	7.742	3.339	1739	160	0.019	0.008	0.030	158	3.49	.001	.052
Controllability [B]	6.400	2.825	1739	160	0.021	0.009	0.034	158	3.33	.001	.052
Awareness	6.853	3.253	1739	160	0.013	0.004	0.022	158	3.00	.003	.054
Awareness [B]	5.537	3.066	1739	160	0.022	0.012	0.032	158	4.19	.001	.052
Correspondence	2.057	0.653	1739	160	0.040	-0.001	0.081	158	1.94	.055	.053
Correspondence [B]	1.608	0.660	1739	160	0.116	0.073	0.158	158	5.38	.001	.047
Concept											
Stereotype	–	–	545	38	0.083	0.042	0.125	37	4.02	.001	.028
Attitude	–	–	1032	119	0.146	0.109	0.182	118	7.77	.001	.065
Identity	–	–	162	11	0.088	-0.090	0.260	10	1.10	.297	.079

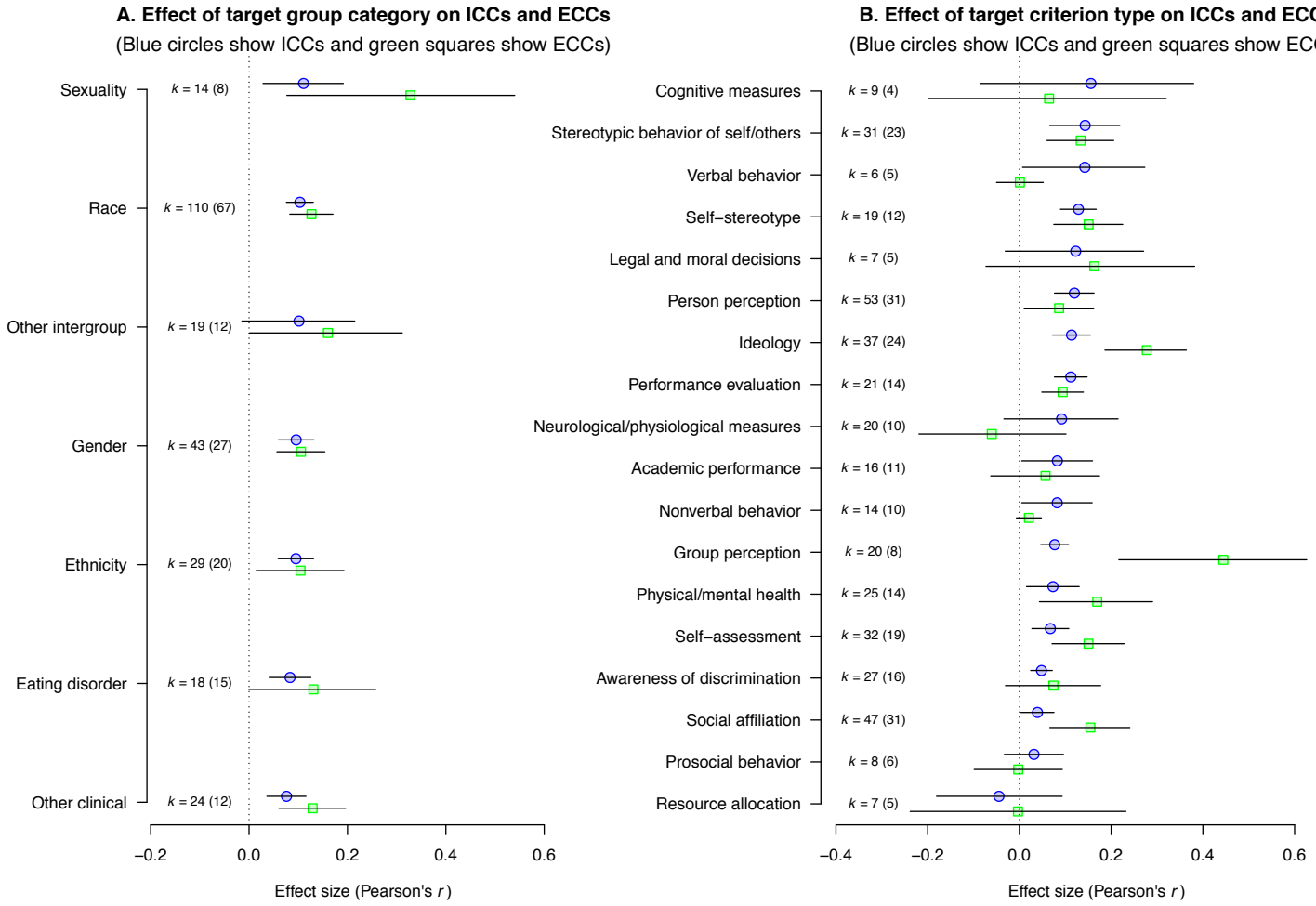

Figure 1. Magnitude of implicit–criterion correlations (ICCs) and explicit–criterion correlations (ECCs) as a function of target group (panel A) and type of target behavior (panel B). The columns on the left display the number of independent effect sizes for ICCs and, in parentheses, for ECCs.