A Meta-Analysis of Procedures to Change Implicit Measures

Patrick S. Forscher
*University of Arkansas, Fayetteville*, forsch@uark.edu

Calvin K. Lai
*Washington University in St. Louis*

Jordan R. Ast
*Duke University*

Charles R. Ebersole
*University of Virginia*

Michelle Herman
*Carolina Outreach, LLC.*

*See next page for additional authors*

Follow this and additional works at: https://scholarworks.uark.edu/psycpub

🔗 Part of the Applied Behavior Analysis Commons, Cognition and Perception Commons, and the Social Psychology Commons

**Recommended Citation**


This Article is brought to you for free and open access by the Psychological Science at ScholarWorks@UARK. It has been accepted for inclusion in Psychological Science Faculty Publications and Presentations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu.
A Meta-Analysis of Procedures to Change Implicit Measures

April 26, 2019

Patrick S. Forscher*¹, Calvin K. Lai*², Jordan R. Axt³, Charles R. Ebersole⁴, Michelle Herman⁵, Patricia G. Devine⁶, and Brian A. Nosek⁴,⁷

*Authors contributed equally to this manuscript and are joint first-authors. Authorship order was determined by coin flip.

¹Dept. of Psychological Science, University of Arkansas.
²Dept. of Psychological & Brain Sciences, Washington University in St. Louis.
³Center for Advanced Hindsight, Duke University.
⁴Dept. of Psychology, University of Virginia.
⁵Carolina Outreach, LLC.
⁶Dept. of Psychology, University of Wisconsin – Madison.
⁷Center for Open Science, Charlottesville, VA.

Paper should be cited as:

In-text citation:
(Forscher, Lai, et al., 2019)

Reference section:

Author notes:
Conceived research: Forscher & Lai, Devine, Nosek; Designed research: Forscher & Lai; Coordinated data collection: Lai; Coded articles: Forscher & Lai, Axt, Ebersole, Herman; Analyzed data: Forscher; Wrote paper: Forscher & Lai; Revised paper: all authors.

We thank Katie Lancaster, Diana Abrego, Amy Bisker, Isabelle Gigante, Julie Lee, Lauren Loffredo, Ryan Massopust, Margot Mellon, Kelci Straka, and Nicole Sather for their assistance with the early phases of article coding. We also thank Mike W.-L. Cheung and Ian White for their assistance and advice about the analyses presented in this paper.

Address correspondence to Calvin K. Lai, Dept. of Psychological & Brain Sciences, Washington University in St. Louis, One Brookings Drive, St. Louis, MO, 63130 (Email: calvinlai@wustl.edu) and/or Patrick S. Forscher, Dept. of Psychological Science, University of Arkansas, 216 Memorial Hall, Fayetteville, AR 72701 (Email: schnarrd@gmail.com).

Data and materials for this project can be found at https://osf.io/awz2p/
Abstract

Using a novel technique known as network meta-analysis, we synthesized evidence from 492 studies (87,418 participants) to investigate the effectiveness of procedures in changing implicit measures, which we define as response biases on implicit tasks. We also evaluated these procedures’ effects on explicit and behavioral measures. We found that implicit measures can be changed, but effects are often relatively weak (|d| < .30). Most studies focused on producing short-term changes with brief, single-session manipulations. Procedures that associate sets of concepts, invoke goals or motivations, or tax mental resources changed implicit measures the most, whereas procedures that induced threat, affirmation, or specific moods/emotions changed implicit measures the least. Bias tests suggested that implicit effects could be inflated relative to their true population values. Procedures changed explicit measures less consistently and to a smaller degree than implicit measures and generally produced trivial changes in behavior. Finally, changes in implicit measures did not mediate changes in explicit measures or behavior. Our findings suggest that changes in implicit measures are possible, but those changes do not necessarily translate into changes in explicit measures or behavior.

Keywords: meta-analysis, implicit measures, implicit bias, intervention, social cognition
A Meta-Analysis of Procedures to Change Implicit Measures

What we intend to do often conflicts with what we actually do. We may plan to diet but find ourselves reaching for a chocolate bar over an apple. We might try to quit smoking but find the temptation of cigarettes too difficult to resist. We may value racial equality but choose to hire a White job candidate over a similarly qualified Black job candidate (Bertrand & Mullainathan, 2004). These gaps between intentions and actions characterize many societal problems, such as intergroup discrimination (Devine, 1989), depression (Beevers, 2005; Haeffel et al., 2007), and addiction (Wiers et al., 2010).

The prevalence of unwanted behaviors across many areas of human life suggests that mental processes outside of one’s conscious awareness or control influence behavior (Smith & DeCoster, 2000). Based on this reasoning, researchers have developed dual-process theories that distinguish between automatic mental processes which are relatively fast, efficient, uncontrollable, and unintentional, and deliberate mental processes which are relatively slow, inefficient, controllable, and intentional. By this logic, the same underlying mental construct can be retrieved either automatically or deliberately. For example, the association between the concepts “Flowers” and “Good” can be retrieved automatically, as when a person spots a vase of flowers and feels good, or deliberately, as when a person thinks about how much they like flowers.

Many dual process theories posit that deliberate processes are more influential on behavior when people have sufficient motivation, awareness, and the ability to reflect before acting, whereas automatic processes are more influential when motivation, awareness, or the ability to reflect are compromised (Devine, 1989; Fazio & Olson, 2014; cf. Greenwald et al., 2009, Kurdi et al., 2018). Many dual process theories also predict that dissociations between intentions and behavior are most likely to occur when the output of automatic and deliberate processes are opposed. Given opposing automatic and deliberate processes, lack of motivation, awareness, or the ability to reflect can cause people to act against their intentions.

Dual process theories are attractive on theoretical and practical grounds. Theoretically, they provide a parsimonious approach for explaining dissociations between intentions and behavior and between mental phenomena more broadly. Dual process theories are used to account for such wide-ranging phenomena as attention (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), reasoning (Evans, 1989; Sloman, 1996; Stanovich & West, 2000), decision-making (Barbey & Sloman, 2007; Kahneman, 2011), memory (Jacoby & Dallas, 1981; Roediger, 1990), attitudes (Wilson, Lindsey, & Schooler, 2000), stereotypes and prejudice (Devine, 1989), the self (Schnabel & Asendorpf, 2010), motivation (Chartrand & Bargh, 2002), and emotion regulation (Mauss, Bunge, & Gross, 2007). Practically, dual-process theories suggest a solution to problems caused by unintentionally biased behavior: change the automatic processes and changes in the behavior influenced by those processes will follow (Forscher & Devine, 2014; Lai, Hoffman, & Nosek, 2013).
Implicit and explicit tasks that assess mental associations between concepts have been a particular interest for dual process theorists. Implicit tasks assess associations through behavior that does not require deliberate retrieval of the target association (e.g., the speed of sorting words into different categories relevant to the association). In contrast, explicit tasks assess associations through behavior that requires deliberate retrieval of the target association (e.g., answers to a questionnaire). For this paper, we define tasks as procedures designed to generate behavioral responses for data analysis. We distinguish tasks from measures, which we define as the outcome of a data-analytic technique applied to behavioral responses (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). On an implicit task, comparisons between responses that result from pairings between one set of concepts relative to responses from a different pairing is referred to as an implicit measure of response bias. Similar comparisons on an explicit task are referred to as an explicit measure of response bias. For example, differences in the time to classify the words “good” or “bad” when they are preceded by the word “flower” or a neutral word can serve as an implicit measure, whereas differences in ratings of the degree to which flowers are good and bad can serve as an explicit measure.

Response biases indexed by implicit and explicit measures are often assumed to reflect automatically or deliberately retrieved associations, respectively. However, like all psychological assessments, implicit and explicit measures are not process-pure. Implicit measures can be influenced by deliberate processes and explicit measures can be influenced by automatic processes (Gawronski & Bodenhausen, 2011). Implicit and explicit measures are also prone to measurement error (e.g., task-switching ability and impulse inhibition for implicit measures, social desirability and acquiescence bias for explicit measures; Blanton et al., 2006; Calanchini et al., 2013; Conrey et al., 2005; Cronbach, 1946; Crowne & Marlowe, 1960).

Implicit and explicit measures are correlated, but the extent to which they correlate varies (Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald et al., 2009; Hofmann et al., 2005; Nosek & Hansen, 2008). These correlations range from very low ($r = .07$; e.g., attitudes toward approaching vs. avoiding) to very high ($r = .70$; e.g., attitudes toward Democrats vs Republicans; Nosek & Hansen, 2008). Half of the variation in implicit-explicit relations can be accounted for with four aspects of the social and mental context: the social sensitivity of the target concepts, the extent to which people have thought about the concepts, the degree to which the concepts in the implicit task are diametrically opposed (e.g., pro-choice vs. pro-life) or not (e.g., dog vs. furniture), and the degree to which people view their opinions about the concepts to be distinct from others (Nosek, 2005; 2007). The predictability of the relation between implicit and explicit measures suggest underlying mental processes that are causally related and/or influenced by third variables (see Fazio & Olson, 2014; Gawronski & Bodenhausen, 2011).

---

1 In the present article, we describe implicit and explicit tasks as assessing an association that is retrieved automatically or deliberately. We are theoretically uncommitted to whether implicit and explicit tasks assess a common representation or categorically different representations, and whether the measures are assessing stored representations or active constructions (Greenwald & Nosek, 2008). Likewise, we use “association” with a theory-uncommitted view (Greenwald et al., 2005). We do not assert a commitment to a particular understanding of what the underlying constructs or processes are (e.g., associative or propositional; Gawronski & Bodenhausen, 2006). Various accounts of the underlying constructs / processes can be adapted to accommodate the changes in implicit measures observed in the present meta-analysis.
Discrepancies between intentions and behavior may arise when automatic and deliberate processes are not aligned, such as intending to be unbiased in selection of candidates for an honor society but showing racial discrimination anyway (Axt, Ebersole, & Nosek, 2014). Consistent with dual process theories, some evidence suggests that implicit measures are more correlated with behavior than explicit measures in socially sensitive issues (Greenwald et al., 2009; cf. Kurdi et al., 2018, Oswald et al., 2013), whereas explicit measures are more correlated with behavior than implicit measures when the situation demands a more deliberate response (Devine, 1989; Fazio & Olson, 2014; Kurdi et al., 2018; cf. Greenwald et al., 2009). Alternatively, when automatic and deliberate processes are aligned, these processes mutually reinforce each other to guide behavior. Supporting this claim, behavior is most consistent with both implicit and explicit measures when implicit and explicit measures are more strongly correlated (Greenwald et al., 2009; Kurdi et al., 2018).

**Change in implicit measures**

Of course, correlation is not causation, so understanding the causal importance of automatically retrieved associations requires procedures that can change automatically retrieved associations. At first, the prospect of changing implicit measures through randomized experiments was dim. Approaches such as cognitive dissonance reduction and persuasive appeals were successful changing self-reported attitudes but often had limited impact on implicit measures (for reviews, see Cooper, 2007; Gawronski & Strack, 2012; Petty & Cacioppo, 1986). The apparent rigidity of automatic processes led the social psychologist John Bargh to portray them as a “cognitive monster” (Bargh, 1999) that is deep-rooted, immune to social pressure, and resistant to the influences of deliberate processes.

Yet this understanding shifted with the discovery that brief experiences can change implicit measures without affecting explicit measures, at least in the short-term (Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000). Over the past sixteen years, the accumulated evidence suggests that implicit measures can be changed, but doing so often relies on mechanisms that are ineffective for shifting explicit measures (for reviews, see Blair, 2002; Dasgupta, 2009; Gawronski & Bodenhausen, 2011; Gawronski & Sritharan, 2010; Lai et al., 2013; Lenton et al., 2009; Sritharan & Gawronski, 2010). For example, the mere presence of a Black experimenter changed implicit measures without affecting explicit measures (Sinclair, Lowery, Hardin, & Colangelo, 2005). More recently, some studies suggest that approaches that affect explicit measures can also affect implicit measures, such as intergroup contact, social threat, and cognitive balance (Bradley et al., 2012; Shook & Fazio, 2008; Smith, De Houwer, & Nosek, 2012). Further, some strategies highlight the process-impurity of implicit tasks by changing aspects of performance in implicit tasks that are unrelated to associative processes (e.g., instruction to fake on an implicit task; Fiedler & Bluemke, 2005; Kim, 2003).

Inspired by social problems characterized by unintentional or unwanted behavior, many studies aim to change automatically retrieved associations with the goal of changing behavior. Many of these studies occur in domains, such as race relations or addiction, where automatic and deliberate processes are often thought to be at odds and where deliberate processes are either resistant to change or theorized to have a limited influence on behavior (e.g., Mann & Kawakami,
2012; Wiers et al, 2010). If intervening on deliberate processes is of limited utility, perhaps intervening on automatically retrieved associations will be more effective.

Despite the proliferation of many approaches to changing implicit measures, little is known about their relative effectiveness (Lai et al., 2013; cf. Lai et al. 2014; 2016). At the same time, there is also little understanding about what approaches are consistently effective across a wide range of phenomena, and what kinds of approaches are inconsistently effective and are contextually dependent on the population, study methodology, or topic of study. Advances in these areas of knowledge would inform a basic understanding of the mental mechanisms that are most influential in changing automatically retrieved associations and a practical understanding of what interventions would be most effective for addressing problems caused by these associations.

Overview of present research

We conducted a meta-analytic review to understand the relative effectiveness of different procedures to change implicit measures and whether changes in implicit measures generalize to changes in explicit and behavioral measures. The diversity in research goals means that research on implicit measure change spans many disciplines, theoretical perspectives, and methodological approaches. Study designs range from two-condition single-session laboratory experiments (e.g., Rudman & Lee, 2002) to multiple-condition longitudinal studies (Sportel, de Hullu, de Jong, & Nauta, 2013). They also differ in what kinds of manipulations are used, from minimal manipulations that prime a concept in memory (Dasgupta & Greenwald, 2001) to intensive long-term interventions that unfold over several weeks (O’Brien et al., 2010). The studies are also diverse in their use of implicit tasks, ranging from popular tasks such as the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) to less popular tasks such as the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010; Hussey et al., 2015).

This research diversity poses two unique analytic issues for meta-analysis. First, different studies often compare different sets of procedures. The diversity in procedures is a challenge for conventional meta-analytic methods that synthesize two-group studies because conventional methods assume all studies use a common comparison. Second, studies in this literature sometimes compare the effects of three or more procedures within the same design. Conventional meta-analytic methods assume that each effect size is independent and thus cannot accommodate these non-independent comparisons.

We imported a technique from the medical sciences called multivariate network meta-analysis to address these issues (Caldwell, Ades, & Higgins, 2005; Lu & Ades, 2004; Salanti, 2012). Compared to conventional meta-analytic methods, network meta-analysis synthesizes information from many procedures simultaneously to better address research literatures where there are many studies that compare distinct procedures (Lumley, 2002). A multivariate implementation of network meta-analysis addresses the problem of single studies making multiple comparisons by modeling the non-independence between multiple comparisons extracted from the same study (White et al., 2012; Mavridis & Salanti, 2012). Multivariate network meta-analysis therefore allows us to use all information from studies comparing many procedures to change implicit measures, rather than having to simplify the information available
when a study has more than one possible contrast (e.g., via averaging, dummy-codes, or data exclusions).

Our meta-analysis was guided by 6 central questions:

1. **What approaches to changing implicit measures are most influential?** We developed a taxonomy of procedures to change implicit measures and compared the effectiveness of procedures within that taxonomy.

2. **Are the sample, methodology, or topic of a study associated with the magnitude of change in implicit measures?** We assessed whether any of these characteristics were associated with the degree of implicit measure change.

3. **How do changes in implicit measures correspond with changes in explicit measures?** We compared the relative size of explicit measure change to implicit measure change. We also examined whether implicit measure change mediated explicit measure change and whether correspondence was larger for studies that used a similar measurement strategy across implicit and explicit tasks.

4. **How do changes in implicit measures correspond with changes in behavior?** We compared the relative size of behavioral change to implicit measure change. We also examined whether implicit measure change mediated behavioral change and whether correspondence was related to the study measurement strategy and the properties of the behavioral task.

5. **Is there evidence that the size of reported effects is biased?** We used three approaches to examine whether reported effect sizes are inflated relative to their true population values and examined three possible mechanisms that might contribute to biased effect sizes (i.e., decline effect, publication bias, United States bias).

6. **Are the results robust to an alternative coding scheme?** We examined whether the conclusions drawn from questions 1-4 were sensitive to an alternative coding scheme focused on the distinction between learning and context (Gawronski et al., 2010; 2015).

**Method**

**Inclusion criteria**

Valid meta-analysis requires careful consideration of which studies are relevant to the research question and which studies are not. We set the following inclusion criteria:

1. **The study is a between-subjects experiment.** We excluded studies that used correlational or quasi-experimental designs (e.g., Rudman, Ashmore, & Gary, 2001) and manipulations that were exclusively within-subjects (e.g., Wheeler & Fiske, 2005). We also excluded studies that experimentally manipulated the stimuli or categories in an
implicit task (e.g., by manipulating whether pictures of animals and plants in an animal/plant pleasant/unpleasant IAT are positively or negatively valenced; Govan & Williams, 2004) because the conditions assessed categorically different associations rather than changing a particular set of associations.

(2) **The study includes an implicit task that is administered after the onset of the experimental procedure.** Implicit tasks were defined as psychological assessments of associations between concepts that do not require the participant to actively bring to mind the target association. This definition included tasks that are both widely used (e.g., the IAT; Greenwald, et al., 1998; Nosek et al., 2007) and less widely used (e.g., Stereotypic Explanatory Bias; Sekaquaptewa et al., 2003). Tasks for which the experimental procedure began during task instructions or practice trials (e.g., Foroni & Mayr, 2005) or for which it extended into the task (e.g., Huntsinger et al., 2010) were also considered eligible.

(3) **The implicit task assesses a pre-existing association.** We defined a “pre-existing association” as an association that either should theoretically be present or has been empirically shown to be present via a demonstration of response bias on an implicit task within the target population. For example, most non-Black people have a response bias indicating they more easily pair Black people with bad and White people with good than the reverse (Nosek, Smyth, et al., 2007), suggesting the presence of a pre-existing Black-bad White-good association. Based on the nature of the pre-existing association, we defined pairings that strengthen (e.g., Black-bad and White-good) and weaken (e.g., Black-good and White-bad) the measured association. Based on this criterion, we excluded studies that formed a new association (e.g., about fictitious people or social groups; McConnell, Rydell, Strain, & Mackie, 2008), studies of ambivalent or unelaborated associations (e.g., Petty et al., 2006), and studies where the mean-level association was theoretically or descriptively neutral.

(4) **Experimental procedures must fit into a single procedure category, and the study must contain procedures from multiple procedure categories** Procedure categories were created iteratively with the goal of capturing the breadth of approaches in the literature. This iterative process meant that the included procedure categories (and studies) changed during the coding process. Procedures that fit into multiple categories or did not fit into any categories were excluded. If a study only had one condition remaining after exclusions, the full study was excluded. For more information about this criterion, see the section labeled “Experimental procedures” below.

(5) **The study is reported in English.** We excluded studies that were not written in English.

### Article retrieval

---

2 In making these decisions, we assumed that people tend to associate positive attributes with both themselves and with their own groups, and that people tend to possess associations that are commonly present in their culture (e.g., Black people with the attribute “musical”). When we could not make a clear determination, we sought data collected from the target population and/or examined whether a pre-existing association was present in a control condition for the study in question.
Our article retrieval procedure was conducted in three phases between September 2012 and July 2015 and again between August 2016 and October 2016. In the first phase (September 2012 to June 2014; August 2016), we retrieved articles that potentially matched our inclusion criteria. We searched PsycINFO and Web of Science using the following search terms: (names of implicit constructs, tasks, and acronyms, e.g., implicit self-esteem*, affect misattribution procedure, GNAT) AND (malleab* OR chang* OR influenc* OR moderat* OR reduc* OR increas* OR shift* OR alter*) AND (1995 TO 2015). We created the list of eligible implicit tasks and acronyms by compiling lists from published reviews of implicit tasks (Nosek, et al., 2011; Gawronski & Payne, 2010), and from discussions among the lead authors (for the full list, see https://osf.io/awz2p/). We supplemented these results with direct requests for relevant studies through email and the Society for Personality and Social Psychology listserv, and an additional 115 articles from an unpublished meta-analysis of the malleability of implicit intergroup bias. Our search procedure resulted in approximately 4,908 articles that potentially matched our inclusion criteria; see Figure 1 and https://osf.io/6ex3n/ for more details.

In the second phase (September 2012 to October 2014; August 2016 to October 2016), trained coders inspected each article and eliminated articles that did not contain a study matching our inclusion criteria. This process thinned our database to 417 articles, 592 studies, and 690 independent samples.

Finally, for any studies that did not report sufficient data to calculate effect sizes and sampling variances and covariances, we sent emails to the corresponding authors requesting the required statistics (November 2014 to July 2015; October 2016). If the authors did not respond, we sent two follow-up reminder emails. If the data required to calculate effect sizes on the implicit task could not be retrieved for a study, we eliminated that study from the meta-analysis. After eliminations, our final sample represented 87,419 participants and included 342 articles, 492 studies, and 571 independent samples.

Article coding

Coders underwent extensive training to reliably apply the coding scheme. We adopted an iterative process to maximize reliability and validity of the coding scheme and to be responsive to the content of the literature (Lipsey & Wilson, 2001). When coders encountered an ambiguity, they added the ambiguity to the agenda for a weekly coding meeting. During these meetings, we discussed each ambiguity until we reached a consensus for resolution. Some ambiguities revealed issues with the coding scheme. In these cases, we revised the coding scheme and rolled out any required coding changes to all other studies. We have made a detailed description of coding scheme and our data and analysis scripts publicly available at https://osf.io/awz2p/. Anyone who is interested can delve into these materials to assess how the results change with different coding decisions.
Figure 1. PRISMA diagram of our data collection process (adapted from Moher et al., 2009).

* This is a conservative estimate of the total number of records, as articles retrieved through direct requests and the unpublished meta-analysis that were excluded from the meta-analysis were not tracked systematically.

** We do not have a complete breakdown of reasons for excluding records. However, we recoded a random 10% ($N = 486$) of the records for reliability coding and provided exclusion reasons for those records. For detailed information about the results of this coding, see [https://osf.io/6ex3n/](https://osf.io/6ex3n/)
We tested the reliability of our coding scheme in multiple waves. In the first wave, we chose a random sample (stratified by topic) of 50 fully coded articles and assigned each of the five coders 10 articles to double-code. Three variables, self-presentation, impulsiveness, and procedures that involve learning and context, were coded after our first wave. For those variables, all studies were double-coded by two independent coders. Another wave assessed the reliability of our effect size extraction procedures with two independent coders who re-coded 28 total samples containing 28 implicit tasks, 21 explicit tasks, and 20 behavioral tasks. A final wave assessed the reliability of our inclusion criteria. We chose a random sample of 10% \((N = 486)\) articles from the PsycINFO/Web of Science database and re-coded whether the studies within each article should be included or not. If a study/article was excluded in this sample, we described the reason(s) for why it was excluded. The result of the reliability coding found near-perfect inter-rater reliability, Cohen’s \(\kappa = .99\). For a detailed report of the reliability coding and the raw results with exclusion reason codes, see the search procedure supplement at https://osf.io/6ex3n/.

**Experimental procedures** (Cohen’s \(\kappa = .71\)). Each experimental procedure was categorized into one of fourteen categories. We developed these categories based on preliminary searches of the literature and prior reviews of malleability and change in implicit measures (e.g., Blair, 2002; Dasgupta, 2009; Gawronski & Bodenhausen, 2011; Gawronski & Sritharan, 2010; Lai et al., 2013; Lenton et al., 2009; Sritharan & Gawronski, 2010) with the goal of capturing the breadth of approaches that researchers have employed. Two of the fourteen categories (physiological deprivation and satiation) were excluded from the final dataset because there were not enough procedures that fit the description (four and two procedures respectively across four papers).

Researchers often disagree about the whether and how a procedure will change implicit measures. To address this issue and maximize agreement between coders, our coding scheme prioritized procedural elements of the study conditions over theoretical expectations regarding the impact of these procedural elements. For example, conditions from two studies that both give participants instructions to show no bias on an IAT would be placed same category, regardless of whether the authors of the studies differ in their predictions as to whether this condition would produce change in IAT scores (e.g., Kim, 2003; Fiedler & Bluemke, 2005). If a given experimental condition fit into multiple coding categories or did not fit into a category clearly, that condition was excluded from the meta-analysis. As shown in Table 1, our final coding scheme included twelve categories:

1. **Strengthen associations directly** (\(k = 127\)) / **Weaken associations directly** (\(k = 154\)). Some efforts to change implicit measures create experiences that directly affirm or counter one’s own biases (e.g., Blair et al., 2001; Dasgupta & Greenwald, 2001). These two categories created pairings of the concepts used in the implicit task to strengthen or weaken the target automatically retrieved association. For example, exposing people to pictures of admired Black people and despised White people in a study assessing associations between Black people/White people and good/bad would go in the “weaken associations directly” category. In contrast, exposing people to admired White people and disliked Black people would go in the “strengthen associations directly” category (Dasgupta & Greenwald, 2001).
(2) Strengthen associations indirectly ($k = 86$) / Weaken associations indirectly ($k = 154$). A related approach to the first category is creating experiences that bring to mind an idea or mindset that will indirectly affirm or counter one’s pre-existing associations (Blair, 2002). These categories were similar to the “strengthen / weaken associations directly” categories except that the procedures did not directly use concepts used in the implicit task. Instead, these procedures attempted to change associations indirectly through the activation of intermediate concepts or mindsets. For example, taking the perspective of a Black person is theorized to create overlap between a person’s self-concept and Black

Table 1. Taxonomy of experimental procedures.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Samples</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weaken associations directly</td>
<td>153</td>
<td>Direct pairing of concepts in implicit measure</td>
<td>Evaluative conditioning (Olson &amp; Fazio, 2006)</td>
</tr>
<tr>
<td>Strengthen associations directly</td>
<td>127</td>
<td></td>
<td>Persuasive argument (Horacio et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Counterstereotypical exemplars (Dasgupta &amp; Greenland, 2001)</td>
</tr>
<tr>
<td>Weaken associations indirectly</td>
<td>154</td>
<td>Activating ideas/mindsets not in implicit measure</td>
<td>Perspective-taking for attitudes (Todd et al., 2011)</td>
</tr>
<tr>
<td>Strengthen associations indirectly</td>
<td>86</td>
<td></td>
<td>Inducing feelings of power (Gutnoe et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Approach/avoid training for attitudes (Kawahami et al., 2008)</td>
</tr>
<tr>
<td>Goals to weaken bias</td>
<td>92</td>
<td>Inducing a goal related to implicit measure</td>
<td>Implementation intentions (Stewart &amp; Payne, 2008)</td>
</tr>
<tr>
<td>Goals to strengthen bias</td>
<td>37</td>
<td></td>
<td>Making anti-prejudiced norms salient (Wyer, 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Subtly priming a goal (Ferguson, 2008)</td>
</tr>
<tr>
<td>Threat</td>
<td>72</td>
<td>Putting one’s identity at risk</td>
<td>Mortality salience (Jong et al., 2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Giving a speech (Rabbit, 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stereotype threat (Frantz et al., 2004)</td>
</tr>
<tr>
<td>Affirmation</td>
<td>23</td>
<td>Maintaining adequacy of one’s identity</td>
<td>Self-affirmation (Radman et al., 2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Group affirmation (Peach et al., 2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Success feedback (Brown, 2010)</td>
</tr>
<tr>
<td>Positive affective state</td>
<td>26</td>
<td>Positive/negative moods or emotions</td>
<td>Listening to happy/sad music (Birch et al., 2008)</td>
</tr>
<tr>
<td>Negative affective state</td>
<td>27</td>
<td></td>
<td>Watching a funny movie (Cain, 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Writing about a disgusting event (Dasgupta et al., 2009)</td>
</tr>
<tr>
<td>Depletion</td>
<td>26</td>
<td>Depletion of mental resources</td>
<td>Cognitive load (Allen et al., 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thought suppression (Hooper et al., 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ego-depletion (Govorun &amp; Payne, 2006)</td>
</tr>
<tr>
<td>Neutral</td>
<td>428</td>
<td>No features relevant to implicit measure</td>
<td>Baseline control condition</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Exposure to unrelated stimuli (Dasgupta &amp; Rivera, 2005)</td>
</tr>
</tbody>
</table>
peop
peop
people (Galinsky & Moskowitz, 2000; Todd et al., 2011). As most people evaluate themselves positively (Taylor & Brown, 1988), linking Black people to the self creates an indirect link between Black people and positivity that changes implicit racial attitudes. Other examples include taking an abstract construals to associate a temptation with negativity (Fujita & Han, 2009) and changing approach/avoid tendencies to change implicit attitudes toward math versus arts (Kawakami, Steele, Cifa, Phillips, & Dovidio, 2008).

(3) Goals to strengthen bias ($k = 37$) / Goals to weaken bias ($k = 92$). Automatically retrieved associations are sensitive to motivations, goals, and habits (e.g., Fishbach, Friedman, & Kruglanski, 2003; Sinclair et al., 2005). Procedures in these categories gave participants goals to respond on an implicit task in ways that either strengthen or weaken the expression of the pre-existing association. These goals could be created directly, such as by instructing participants to appear non-shy on an implicit task assessing shy/non-shy self-concept (Asendorpf, Banse, & Mucke, 2002). These goals could also created indirectly, such as by making anti-prejudiced norms salient prior to an implicit task assessing attitudes toward Black people (Wyer, 2010).

(4) Threat ($k = 72$). Threat involves putting the integrity of a person’s identity at risk. Threat plays a powerful role in shifting attention (Mogg, Bradley, De Bono, & Painter, 1997), evaluations of one’s self (Taylor & Lobel, 1989), and evaluations of others (Stephan & Stephan, 2000). The threats included in this category were diverse, including the threat of confirming a negative stereotype (e.g., Frantz et al., 2004), mortality salience (e.g., Jong, Halberstadt, & Bluemke, 2012), and the threat of giving a speech in front of a panel of judges (e.g., Rabbitt, 2012).

(5) Affirmation ($k = 23$). Affirmation involves procedures that sought to maintain the adequacy of a person’s identity, which may buffer against acute or chronic experiences of threat (Cohen & Sherman, 2014; Steele, 1988). Examples in this category include procedures in which the participants were given feedback that they were competent, moral, or unbiased (Frantz et al., 2004), and procedures where the participants were instructed to think about a value important to a social group to which they belonged (Peach et al., 2011).

(6) Positive affective state ($k = 26$) / Negative affective state ($k = 27$). According to an affect-as-information account, positive affect affirms chronically accessible concepts and negative affect rejects them (Huntsinger, Isbell, & Clore, 2014). These categories involved procedures that induced a mood or emotion without placing the manipulation in the “threat” or “affirmation” categories. Although manipulations that threaten or affirm a person’s identities are likely to induce affect, we reasoned that threat and affirmation are the primary characteristics of these conditions and take precedence.\(^3\) Examples of manipulations in these categories included both positive or negative mood inductions (e.g., Birch et al., 2008) and inductions of specific emotions like anger, disgust, or moral

\(^3\) We placed anger inductions into the positive affective state category as anger is more cognitively and neurally similar to positive emotions than negative ones (Carver & Harmon-Jones, 2009; Harmon-Jones, 2003; Lerner & Tiedens, 2006).

(7) Depletion ($k = 26$). Depleting mental resources may lead to increased reliance on social-cognitive biases (e.g., Bodenhausen, 1990; Gilbert & Hixon, 1991; Stangor & Duan, 1991). This category involves manipulations that reduced the amount of mental resources available to the participant during the implicit task. Oftentimes, participants were instructed to complete a mentally effortful task prior to or during the implicit task, such as holding a multi-digit number in their heads (Allen et al., 2009).

(8) Neutral ($k = 428$). This category involves conditions where nothing happened that could plausibly affect response biases on implicit tasks (e.g., control conditions). This category did not contain every procedure that a specific research tradition would deem ineffective. For example, on the basis of past evidence (Dasgupta & Greenwald, 2001), some researchers would predict that exposure to images of admired White people and disliked Black people does little to change implicit racial attitudes because admired White people are already chronically accessible. Although this may be the case, exposure to admired White people pairs White people with positivity, and thus this procedure would be placed in the “strengthen associations directly” category.

Implicit, explicit, and behavioral tasks. Tasks were considered implicit if they did not require the target association to be actively brought to mind. For example, the Black/White good/bad IAT requires participants to categorize Black faces, White faces, positive words, and negative words, but it does not require them to introspect about their feelings about Black people relative to White people. Tasks were considered explicit if they required the target association to be actively brought to mind. For example, a survey item asking “How warm do you feel toward Black people?” requires participants to actively assess their personal feelings about Black people. Tasks were considered behavioral if they involved the participant’s actual, hypothetical, or intended behavior in relation to the target association. Behavioral tasks involved a wide range of outcomes, such as seating distance from a Black or White confederate (Todd et al., 2011), willingness to participate in a hypothetical beer pong game (Goodall & Slater, 2010), intentions to drink in the future (Glock, Klapproth, & Müller, 2015; Lindgren et al., 2015), reported chocolate consumption (Kroese, Adriaanse, Evers, & De Ridder, 2011), and intentions to vote for gay and lesbian civil rights referenda (Dasgupta & Rivera, 2008).

Explicit and behavioral tasks were included only if coders judged that they assessed the same association as the implicit task selected from the study. For example, a questionnaire assessing Black stereotypes would be eligible for an implicit task assessing Black/White stereotyping but not an implicit task assessing Black/White attitudes. This inclusion criterion was notably stricter than past meta-analyses that included explicit/behavioral tasks which did not narrowly tap into the same constructs (e.g., physiological or neural activity for IATs in Greenwald et al., 2009; Kurdi et al., 2018; stereotype tasks for attitude IATs in Oswald et al., 2013). As with the implicit tasks, explicit and behavioral tasks were only eligible if they were administered after the onset of the manipulation. If multiple tasks in a sample met our definition of an implicit, explicit, or behavioral task, we selected the task that was most widely used in the meta-analysis (i.e., if a study included both an IAT and a Lexical Decision Task assessing
implicit self-esteem, we selected the IAT) or the task that best matched the implicit task conceptually (e.g., for a relative implicit stereotyping task, we prioritized relative explicit stereotyping tasks over absolute stereotyping tasks).

All measures were scored such that higher numbers represent greater levels of the pre-existing response bias. Implicit tasks that assessed associations between two sets of concepts were scored by creating a difference score that reflected the underlying association. For example, in a study where researchers measured participant reaction times (RT) to categorize positively and negatively valenced words with Black and White face primes, we created the following difference score: (Black/good RT - Black/bad RT) - (White/good RT - White/bad RT). If a score computed from a D score algorithm (Greenwald et al., 2003) was used, we chose that over a reaction time difference score (Nosek & Sriram, 2007). If the explicit and behavioral tasks were composed of multiple parts (e.g., separate assessments of feelings of warmth toward Black people and White people), we scored the aspects of those tasks that were most correspondent with the implicit task. In a study using the aforementioned priming task that also contained separate feelings thermometer ratings of Black people and White people, we created the following difference score: White thermometer rating - Black thermometer rating.

Multiple study subsamples. If a study reported their results separately for groups with a given individual-difference characteristic (e.g., a median split of a questionnaire task), we collapsed across the target individual difference. If, however, participants were recruited on the basis of that individual difference characteristic (e.g., from the top and bottom quartile of a scale), we treated these groups as separate subsamples for the purposes of the meta-analysis to avoid confounding (Glass, 1977). In some cases, we analyzed groups separately even if they were not recruited on a specific characteristic if the meaning of the task or manipulation was unambiguously different for different subgroups. For example, the meaning of a Bill Clinton/George Bush good/bad IAT is likely different for Democrats and Republicans because Democrats share a party affiliation with Bill Clinton, whereas Republicans share a party affiliation with George Bush (Albertson, 2011). Finally, studies were split into subsamples if the study randomly assigned participants to different implicit tasks in addition to randomly assigning them to different manipulations (e.g., by assessing the effects of reading a counter-stereotypical vs. neutral scenario on the personalized vs. original IAT, Han et al., 2010).

Sample characteristics

Sample population (κ = .92). University student samples tend to be more compliant and more easily socially influenced (Sears, 1986), and may be more susceptible to psychological manipulations than non-student samples (e.g., Lai et al., 2016). Student and non-student samples may also differ because of issues related to the publication process (e.g., reviewers may be less critical of small effects if the study does not use an undergraduate convenience sample). To assess these possibilities, we coded whether the sample was drawn from a university student or a non-university-student population (e.g., hazardous drinkers, elementary school children).

Demographic characteristics (% women ɑ = .89; % White ɑ = .96). We coded the racial and gender distribution of each sample to examine the generalizability of results to different demographic groups. Coders recorded the number of participants who were White, non-White,
or whose race was not reported. Coders followed a similar process for gender distribution: male, female, or gender not reported. For analysis, we used the percentage of women and White people in samples that reported that information.

**Methodological characteristics**

**Design** (implicit $\kappa = .86$; explicit $\kappa = .89$; behavior $\kappa = .96$). The effects of procedures on implicit measures may depend on whether participants completed an implicit task before the intervention (e.g., Lai et al., 2014). Thus, we assessed whether implicit, explicit, and behavioral tasks were administered in a fully between-subjects design or in a mixed design with between-subjects and within-subjects (i.e., pre-test and post-test) components.

**Implicit task** ($\kappa = .90$). Different implicit tasks may tap different constructs. Implicit tasks also vary in measurement reliability, which can depress the relationship between manipulations and their effects on implicit measures. To examine these possibilities, we coded the specific implicit task used for each study (e.g., the Affect Misattribution Procedure, Go/No-Go Association Task, Evaluative Priming). As there were not enough studies to test for more nuanced differences, we analyzed data by whether the study’s implicit task was an IAT or not.

**Longitudinal** ($\kappa = .87$). This variable assessed whether the implicit task was administered longitudinally (i.e., at least one of the assessments occurred after a delay that is longer than one experimental session). As only 38 (6.6%) of 598 samples were longitudinal, we did not use this variable for inferential analyses.

**Manipulation length** ($\kappa = .64$). This variable assessed whether the manipulation occurred in a single experimental session or in multiple sessions. Only 17 (3.0%) of the 598 samples had procedures occurring over multiple sessions, so we did not use this variable for inferential analyses.

**Characteristics of explicit and behavioral measures**

**Correspondence between implicit tasks and explicit** ($\kappa = .70$) and **behavioral** ($\kappa = .98$) tasks. The principle of correspondence predicts that measures are better predictors of behavior when they are measured at the same level of specificity (Ajzen, 1988; Fishbein & Ajzen, 1975; Sutton, 1998) and assess the same contents (Gawronski, in press). Supporting this principle, implicit and explicit measures are more strongly correlated with each other when they share the same level of specificity (Axt, 2018; Greenwald et al., 2009; Hofmann et al., 2005). Implicit and behavioral measures are also more strongly correlated with each other when the measures are correspondent, although investigators find this pattern less reliably (Greenwald et al., 2009; Kurdi et al., 2018; Oswald et al., 2013).

There are many approaches to operationalizing correspondence. We examined one such approach: whether measures were assessed using an absolute scale (i.e., a single target, e.g., Flower) or a relative scale (i.e., comparisons between multiple targets, e.g., Flowers vs. Insects). We coded whether implicit and explicit/behavioral measures were both assessed on an absolute scale, a relative scale, or whether one was assessed on an absolute scale whereas the other was
not. For analysis, we compared studies by whether the implicit and explicit/behavioral tasks matched (higher correspondence) or not (lower correspondence).

Degree of impulsiveness/deliberation in the behavior (κ = .83). The MODE model (Fazio, 1990; Fazio & Towles-Schwen, 1999; Olson & Fazio, 2009) predicts that automatically retrieved associations are especially likely to influence behavior when the motivation or opportunity to engage in deliberate mental processing is limited. We coded whether the behavioral task was clearly deliberate (math test performance; Galdi, Cadinu, & Tomasetto, 2014), clearly impulsive (how closely spider-phobics dare to approach a medium-sized house spider; Huijding & de Jong, 2007), or not clearly deliberate or impulsive (amount of time spent reading information about smoking cessation; Macy et al., 2015). To retain statistical power for moderator analyses, we split this three-level variable into two dummy-coded variables that compare one level against the other two (deliberate vs. non-deliberate; impulsive vs. non-impulsive).

Topic characteristics

Evaluative vs. conceptual associations (κ = .85). Implicit associations vary in whether their content is more evaluatively (e.g., good/bad) or conceptually (e.g., masculine/feminine) focused. Because some evidence has suggested that different neural substrates are associated with affective and semantic memory (Amodio, 2018; Amodio & Devine, 2006; Amodio & Ratner, 2011), it is possible that the same procedure will produce different effects on conceptual and evaluative associations. We therefore coded whether the concepts involved in the target association were primarily evaluative (e.g., good/bad in a self/other-good/bad IAT) or conceptual (e.g., science/humanities in a male/female-science/humanities IAT). Some associations had both evaluative and conceptual content (e.g., a Lexical Decision Task where the primes are pictures of Black people and the targets are negative Black stereotypes). We handled these on a case-by-case basis.

Self-associations (κ = .85). The self is one of the most fundamental constructs in psychology (James, 1890), and has long been an important construct in research on automatic processes (Greenwald & Banaji, 1995). Whether self-associations should be more or less easy to change than other associations is unclear. To assess the role of the self in implicit malleability, we coded whether or not the concepts involved in the target association were related to the self.

Association domain (κ = .97). The topics of study in the meta-analysis were diverse, ranging from anti-Arab/Muslim prejudice to dieting and exercise. Coders judged whether the study’s topic was related to intergroup relations, health psychology, personality, clinical psychology, political preferences, consumer preferences, or close relationships. For analysis, we treated this as two separate variables, one that compared intergroup and non-intergroup studies and a second that compared health/clinical studies and non-health/clinical studies.

---

4 We also attempted to code the degree to which the behavioral task invoked self-presentation concerns. However, we were unable to attain acceptable levels of agreement among coders. For more information, see the supplement at https://osf.io/awz2p/.
Article characteristics

*Publication status* ($\kappa = 1.00$). Larger significant effects are more likely to be published than smaller non-significant effects (Stern & Simes, 1997). We assessed whether this was the case in this literature by coding whether a study had been published in an academic journal or book at the time of analysis. Many of the unpublished studies were dissertations and/or studies in a researcher’s “file-drawer,” but some unpublished studies were studies that were in the process of being prepared for publication.

*Publication year* ($\alpha = 1.00$). The effect size of early published studies is often larger than effect sizes of later published studies on the same topic (Jennions & Møller, 2002), a result popularly known as the decline effect. There are multiple possible reasons for the decline effect, including publication bias, increasing sample heterogeneity, and loss of adherence to intervention quality over time. We coded the year a study was published to see if a decline effect exists in this literature. Unpublished studies were not included in any analyses involving publication year.

Study characteristics

*Geographic region of sample* ($\kappa = .92$). Published effect sizes from the United States in the behavioral sciences tend to be larger than those published in other countries, perhaps due to publication pressures (Fanelli & Ioannidis, 2013). To investigate whether this was the case in this literature, we coded whether the studies were conducted in the United States, Europe, Israel, Canada, Australia and New Zealand, Asia, Africa and Latin America, or multiple countries. For analysis, we compared the effect sizes of studies published in the US and elsewhere.

*Number of experimental groups* ($\kappa = .67$). This variable represented the number of groups in the study’s design, as determined by the study’s author. Sometimes this variable was synonymous with the number of conditions we used in analysis, but often times it was not (e.g., when a condition was excluded, when multiple conditions were merged together for analysis). For moderator analysis, we compared studies that used a two-group design to studies that had more than two groups.

Meta-analytic computations

Meta-analysis involves the synthesis of one or more effect sizes and the sampling variances associated with those effect sizes. The breadth of this project demanded special procedures to do so.

*Standardized mean differences* ($a_{\text{implicit ES}} = 1.00; a_{\text{explicit ES}} = 1.00; a_{\text{behavior ES}} = .97;^5 a_{\text{implicit var.}} = 1.00; a_{\text{explicit var.}} = 1.00; a_{\text{behavior var.}} = 1.00$). Differences between groups were assessed using the standardized mean difference. For each comparison between procedures, we estimated Hedge’s $g$ (Hedges & Olkin, 1985), which is a measure similar to Cohen’s $d$ that corrects for small-sample bias. We estimated Hedge’s $g$ using the raw (non-covariate-adjusted) means,

---

5 This excludes a single study error in which the effect size for a study with $N = 109$ was coded in the wrong direction. When this single study is included in the $\alpha$ calculation, the behavioral Krippendorff’s $\alpha = .60$. 
standard deviations, and number of participants within each cell of a given sample’s design. To calculate the pooled standard deviation for the Hedge’s $g$ denominator, we pooled the standard deviations across all cells of a given sample’s design. If the total sample size was available but the number of participants per group was not, we assumed equal sample sizes within each group. If the means and/or standard deviations were missing, we attempted to back-calculate the missing descriptive statistics or the standardized mean difference from other statistics reported in the article (see Lipsey & Wilson, 2001). If this was not possible, we requested the required information directly from the authors.

In multi-group designs (i.e., designs with more than two groups), we designated one group the “reference group” and computed multiple effect sizes relative to this reference group (Salanti, 2012; White et al., 2012). This yielded $(g - 1)$ effect sizes, where $g$ is the number of groups in a study. Where possible, this reference group was a neutral condition. In studies that lacked a neutral condition, we calculated effect sizes relative to a virtual neutral condition that had an effect size of 0 and a standard error of 1000 (Higgins & Whitehead, 1996; White et al., 2012). This computational device ensures that studies that lack a neutral condition will contribute information during model fitting (Higgins & Whitehead, 1996) without directly influencing meta-analytic estimates involving neutral conditions (White et al., 2012). The virtual neutral conditions therefore play a similar role as continuity corrections to avoid divide-by-zero errors when analyzing odds ratios: they allow estimation to proceed without inappropriately impacting results.

We handled experiments with pre-test post-test designs by using the mean differences from pre-test to post-test as the means within each condition and the pre-test standard deviations as our standard deviations within each condition (Morris & DeShon, 2002; Morris, 2008). If the pre-test standard deviations were unavailable but the standard deviations of the differences from pre-test to post-test were available, we used the standard deviations of the differences instead, then transformed this change score metric into one comparable to the pre-test standard deviation metric (Morris & DeShon, 2002). If we were unable to obtain either the pre-test or difference score data, we computed effect sizes with post-test data only. Some studies used dichotomous outcomes to assess behavior. For these outcomes, we calculated log-odds ratios that we then translated into a metric equivalent to standardized mean differences (Cox & Snell, 1989; Sánchez-Meca et al., 2003).

**Sampling variances and covariances.** The sampling variances of Hedge’s $g$ in post-test only designs were estimated using formulas developed by Hedges and Olkin (1985). In experiments with pre-test post-test designs, we estimated the variances using formulas that correct for the correlation between pre-test and post-test (Morris & DeShon, 2002; Morris, 2008). For studies missing the correlation between pre-test and post-test (27/84 implicit correlations; 11/35 explicit correlations, 3/14 behavioral correlations), we imputed the missing correlation with its meta-analytic estimate calculated from the rest of the sample (implicit $r = .35, k = 57, 95\% \text{ CI} = [.29, .41]$; explicit $r = .74, k = 24, 95\% \text{ CI} = [.68, .79]$; behavioral $r = .72, k = 11, 95\% \text{ CI} = [.66, .78]$). We estimated the variances for effect sizes of dichotomous tasks using a formula described by Cox and Snell (1989).
Effect sizes extracted from a single study are typically non-independent, either because they share a common reference group in multi-group studies or because the same participants complete multiple tasks (i.e., when participants take an implicit task and an explicit or behavioral task). Thus, in addition to the variances typically estimated in pairwise meta-analyses, we also estimated covariances between each pair of effect sizes derived from a given study in studies that yielded multiple effect sizes. For multi-group studies, estimating the covariance between effect sizes only requires the number of people per condition and the means and standard deviations of the outcome measure (Gleser & Olkin, 2009). For studies with multiple measures (i.e., an explicit and/or behavioral measure in addition to an implicit measure), the calculation of these covariances requires the correlation between the two types of measures. In studies where this correlation was unavailable (26/260 implicit-explicit correlations; 12/94 implicit-behavioral correlations), we imputed the correlation using the meta-analytic estimate from the remaining studies (implicit-explicit $r = .14, k = 228, 95\% CI = [.12, .16]$; implicit-behavioral $r = .09, k = 80, 95\% CI = [.07, .14]$). We estimated the covariances between different measures using formulas derived by Wei and Higgins (2013).

**Indirect effects.** We computed indirect effects to estimate the degree to which the effects of procedures on explicit or behavioral measures was mediated by change in implicit measures. To obtain these estimates, we constructed a series of 3 by 3 correlation matrices representing the bivariate relationships between manipulations, implicit measures, and explicit/behavioral measures. The correlations between manipulations and other variables were extracted for each study report by transforming the standardized mean differences on implicit measures and explicit/behavioral measures into correlation coefficients. These correlations were combined with the correlation between implicit measures and explicit/behavioral measures. We only included two-condition studies when constructing these correlation matrices because of ambiguity in how to define the direct and indirect effects in multi-condition studies. We then used the delta method to extract the standardized indirect effects and their asymptotic variances from these correlation matrices (Cheung, 2009).

**Results**

**Network meta-analysis**

We performed most of the analyses using a multivariate implementation of network meta-analysis (Caldwell et al., 2005; Lu & Ades, 2004; Salanti, 2012). Multivariate network meta-analysis treats each study in the meta-analysis as having multiple potential outcomes. Each of these outcomes is a potential comparison between 2 of the 12 categories of procedures coded for the meta-analysis. Comparisons that are not present in a given study are treated as missing values. For example, a two-group study would have one comparison and many missing values for all the comparisons that were not tested. Because studies that contain more than two categories of procedures yield more than one two-group comparison, multivariate network meta-analysis explicitly models the interdependence between these multiple comparisons.

---

6 Although we imputed this correlation for the analysis of the consistency between effects on implicit measures and explicit/behavioral measures, we did not impute this correlation for the analysis of the indirect effects.

7 We also estimated the direct effects, their asymptotic variance, and the asymptotic covariance between the direct and indirect effect so as to not bias the indirect effect estimates. We only report the indirect effects here.
More formally, given \( k \) studies comparing \( g \) conditions, multivariate network meta-analysis represents each study as a set of comparisons between one of the conditions (the reference group \( r \)) and each other condition. Thus, study \( i \) yields a vector of \((g - 1)\) effect sizes, labeled \( y_i \), along with a \((g - 1) \times (g - 1)\) matrix of variances and covariances between the effect sizes within study \( i \), labeled \( S_i \). Given effect sizes \( y_i \) and covariance matrices \( S_i \), one can estimate coefficients \( \alpha \) and the between-studies variance-covariance matrix \( \Sigma \) using the following multivariate model (White et al., 2012):

\[
y_i \sim N(\alpha X_i, \Sigma + S_i)
\]

where \( X_i \) is a matrix of study covariates. If there are no study covariates and \( \alpha \) and \( \Sigma \) are assumed to be the same across studies, \( \alpha \) represents the meta-analytic effect size estimates of comparisons between the reference group and each other condition and \( \Sigma \) represents the between-studies variance-covariance matrix for those effect sizes.

An advantage of this meta-analytic model is that it uses both direct information from the comparisons within each study and indirect information from the pattern of comparisons across studies (Higgins & Whitehead, 1996; Lu & Ades, 2004). For example, taking the difference between the effect of the comparisons between procedures A & B and procedures A & C allows for the indirect estimation of the comparison of procedures B & C. Direct and indirect information can only be combined if a network of comparisons meets the consistency assumption, which assumes that each procedure is similar regardless of which other procedures appear alongside it in a given study (Salanti, 2012). We tested the viability of this assumption by testing whether, within single treatment estimates, studies of different designs had different effect sizes (the design by treatment interaction approach; White et al., 2012; White, 2015). They did not, \( \chi^2(71, k = 571) = 86.11, p = .107 \), indicating the consistency assumption was reasonable for our data.

We fit all multivariate network meta-analytic models using the metaSEM package in R (Cheung, 2015). To ensure model identifiability, we constrained the components of the between-studies variance-covariance matrix \( \Sigma \) such that the variances were equal and the covariances were equal (Higgins & Whitehead, 1996; Lu & Ades, 2004).  

---

\(^8\) We explored the viability of a model that allows the variances to be unequal but still constrains the covariances to be equal. This model had better fit than the more constrained model, \( \chi^2(10, k = 571) = 32.12, p < .001 \). However, as we show in our supplement at https://osf.io/ejzf7/, the estimated effects of the procedures on implicit measures were highly similar across the constrained and less constrained models, and the less constrained model had issues with model identifiability when we attempted to fit more complicated models than the one with just the implicit effects (for example, moderator models or models of the correspondence between implicit and explicit effects). For these reasons, we present the models with the more constrained variance-covariance matrix throughout the text.
Table 2. Characteristics of the final meta-analysis sample.

<table>
<thead>
<tr>
<th>Methodological characteristics</th>
<th>Sample characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proc. length</strong></td>
<td><strong>Population</strong></td>
</tr>
<tr>
<td>Single session</td>
<td>University student 467 81.8%</td>
</tr>
<tr>
<td>Multiple sessions</td>
<td>Not university student 104 18.2%</td>
</tr>
<tr>
<td><strong>Longitudinal</strong></td>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Longitudinal</td>
<td>Female 52,345 65.6%</td>
</tr>
<tr>
<td>Non-longitudinal</td>
<td>Male 27,442 34.4%</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td><strong>Race</strong></td>
</tr>
<tr>
<td>Post-test only</td>
<td>Not reported 15,199</td>
</tr>
<tr>
<td>Pre-test post-test</td>
<td>White 42,403 76.2%</td>
</tr>
<tr>
<td><strong>Implicit task</strong></td>
<td><strong>Location</strong></td>
</tr>
<tr>
<td>IAT</td>
<td>Non-White 13,272 23.8%</td>
</tr>
<tr>
<td>Priming</td>
<td>Not reported 39,082</td>
</tr>
<tr>
<td>SC-IAT/ST-IAT</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td><strong>Explicit task</strong></td>
<td><strong>Conditions</strong></td>
</tr>
<tr>
<td>Present</td>
<td>Two 236 48.0%</td>
</tr>
<tr>
<td>Not present</td>
<td>Three 89 18.1%</td>
</tr>
<tr>
<td><strong>Behavioral task</strong></td>
<td><strong>Four</strong></td>
</tr>
<tr>
<td>Present</td>
<td>102 20.7%</td>
</tr>
<tr>
<td>Not present</td>
<td></td>
</tr>
<tr>
<td><strong>Explicit / behavioral characteristics</strong></td>
<td><strong>Five+</strong></td>
</tr>
<tr>
<td>I/E corresp. Higher</td>
<td>65 13.2%</td>
</tr>
<tr>
<td>Lower</td>
<td></td>
</tr>
<tr>
<td>I/B corresp. Higher</td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td></td>
</tr>
<tr>
<td><strong>B deliberation</strong></td>
<td><strong>Status</strong></td>
</tr>
<tr>
<td>Impulsive</td>
<td>Published 277 80.8%</td>
</tr>
<tr>
<td>Deliberate</td>
<td>Unpublished 66 19.2%</td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
</tr>
<tr>
<td><strong>Topic characteristics</strong></td>
<td><strong>Date</strong></td>
</tr>
<tr>
<td>Domain</td>
<td>1995 - 2000 3 1.1%</td>
</tr>
<tr>
<td>Intergroup</td>
<td>2001 - 2005 31 11.2%</td>
</tr>
<tr>
<td>Personality</td>
<td>2006 - 2010 87 31.4%</td>
</tr>
<tr>
<td>Health/clinical</td>
<td>2011+ 156 56.3%</td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Evaluative</td>
<td></td>
</tr>
<tr>
<td>Conceptual</td>
<td></td>
</tr>
<tr>
<td>Self-related Non-self</td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td></td>
</tr>
</tbody>
</table>

Note. Methodological, topic, and sample characteristics are presented in # of samples. Gender/Race are presented in # of participants. Study characteristics are presented in # of studies. Publication status is presented in # of papers, and publication date is presented in # of published papers.
Descriptive information about the articles, studies, samples, and tasks included in the meta-analysis is shown in Table 2. The data primarily came from published articles (80.8%), studies conducted in the United States (53.0%), and from studies of intergroup relations (63.5%). The participants in the meta-analysis reflect the demographics of students in Introductory Psychology classes: 81.8% of samples were composed entirely of university students, and samples were majority White (76.2%) and female (65.6%). The majority of the samples used evaluative tasks (65.0%), usually with an IAT (64.8%), and usually in a single-session, post-test only design (83.9%). Only 38 (6.7%) of the samples used a longitudinal design to assess change over time, and only 17 (3.0%) used intense, multi-session procedures. Finally, 45.5% of the samples included an explicit task, and 16.5% of the samples contained a behavioral task.

Most study characteristics were weakly correlated. Some of the strongest relationships involved health/clinical studies. Compared to studies in other domains, health/clinical studies were more likely to use a pre-test post-test design (r = .41) and include a behavioral task (r = .38). When health/clinical studies used a behavioral task, the task was also less likely to be categorized as deliberate (r = -.43). For a complete correlation matrix of study characteristics, see https://osf.io/awz2p/.

The network of comparisons between the 12 categories of procedures is shown in Figure 2. The most common procedure most frequently used in a study was the neutral category. Indeed, most studies (75.0%) compared neutral procedures with one or more comparison procedures. When studies made other types of comparisons, they most often (86.7%) compared a procedure and its conceptual opposite (e.g., positive and negative affective states). Few studies that made non-neutral comparisons used procedures in conceptually different categories (13.2%) (e.g., weaken associations directly vs. threat).

What approaches to changing implicit measures are most influential?

We compared the effectiveness of procedures to change implicit measures by fitting a multivariate network meta-analytic model with the neutral group as the reference category. As shown in Figure 3, seven categories changed implicit measures relative to a neutral condition: procedures that strengthen or weaken associations, either directly (g_strengthen = .21, 95% CI = [.13, .28]; g_weaken = -.23, 95% CI = [-.30, -.16]) or indirectly (g_strengthen = .14, 95% CI = [.04, .24]; g_weaken = -.23, 95% CI = [-.30, -.16]), that induce goals (g_strengthen = .14, 95% CI = [.00, .28]; g_weaken = -.29, 95% CI = [-.37, -.21]), and that deplete mental resources (g = .24, 95% CI = [.07, .40]). In all cases, the average effects were small by conventional standards (|d| < .35; Hyde, 2005) and below the median effects reported in social psychology papers (median d = .37; Richard, Bond, & Stokes-Zoota, 2003). Compared to a neutral procedure, procedures that produce threat (g = .08, 95% CI = [-.02, .18]), affirmation (g = -.02, 95% CI = [-.20, .17]), positive affective states (g = -.06, 95% CI = [-.24, .11]), and negative affective states (g = -.12, 95% CI = [-.31, .07]) produced effects that were small and not distinguishable from zero.

We estimated the variation in effect sizes due to substantive differences between studies using the multivariate R-based statistic developed by Jackson, White, and Riley (2011). This
statistic revealed high between-study variation, at least as compared to the typical study sampling variance ($I^2 = .809$), a finding mirrored by the large estimated effect size standard deviation ($\tau = .306$). This reflects the diversity of disciplines, theoretical approaches, and methodological approaches in this area.

Figure 2. Network plot of procedures included in the meta-analysis. The radius of the category circles = the number of procedures in that category, line width = the number of samples in which a pair of conditions were directly compared.
Figure 3. Forest plot of the comparisons between each procedure and a neutral procedure. $k$ gives the number of studies that directly (or indirectly, listed in parentheses) compare the listed procedure and a neutral procedure. $g$ gives the estimated standardized mean difference and its 95% CI. Higher effect sizes reflect greater increases in the implicit measure relative to a neutral procedure.

<table>
<thead>
<tr>
<th>Procedure category</th>
<th>$k$</th>
<th>$g$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weaken associations directly</td>
<td>104 (70)</td>
<td>-0.23 [-0.30, -0.16]</td>
</tr>
<tr>
<td>Strengthen associations directly</td>
<td>79 (48)</td>
<td>0.21 [0.13, 0.28]</td>
</tr>
<tr>
<td>Weaken associations indirectly</td>
<td>105 (58)</td>
<td>-0.23 [-0.30, -0.16]</td>
</tr>
<tr>
<td>Strengthen associations indirectly</td>
<td>40 (46)</td>
<td>0.14 [0.04, 0.24]</td>
</tr>
<tr>
<td>Goals to weaken bias</td>
<td>76 (16)</td>
<td>-0.29 [-0.37, -0.21]</td>
</tr>
<tr>
<td>Goals to strengthen bias</td>
<td>22 (15)</td>
<td>0.14 [0.00, 0.28]</td>
</tr>
<tr>
<td>Threat</td>
<td>53 (19)</td>
<td>0.06 [0.02, 0.18]</td>
</tr>
<tr>
<td>Affirmation</td>
<td>10 (13)</td>
<td>-0.01 [-0.20, 0.17]</td>
</tr>
<tr>
<td>Positive affective state</td>
<td>12 (14)</td>
<td>-0.06 [-0.24, 0.11]</td>
</tr>
<tr>
<td>Negative affective state</td>
<td>8 (19)</td>
<td>-0.12 [-0.31, 0.07]</td>
</tr>
<tr>
<td>Depletion</td>
<td>24 (2)</td>
<td>0.24 [0.07, 0.40]</td>
</tr>
</tbody>
</table>

Are the sample, methodology, or topic of a study associated with the magnitude of implicit change in implicit measures?

We tested whether effect sizes varied according to the sample, design, or topic of a study. We did this by using Wald $\chi^2$ tests that compared moderator models to models without any moderators. There was evidence of variation based on whether the sample was a student sample, $\chi^2(10, k = 571) = 26.34, p = .003$, the racial composition of the sample, $\chi^2(11, k = 247) = 20.50, p = .039$, the implicit task, $\chi^2(10, k = 571) = 32.12, p < .001$, whether the design included a pre-test implicit task, $\chi^2(11, k = 571) = 37.16, p < .001$, and whether the target association was related to the self, $\chi^2(11, k = 571) = 22.75, p = .019$. There was little evidence of variation by the number of conditions compared within the study, $\chi^2(11, k = 571) = 13.04, p = .291$, the gender composition of the sample, $\chi^2(11, k = 482) = 14.85, p = .189$, the target association was evaluative or conceptual, $\chi^2(11, k = 571) = 19.08, p = .060$, whether the target association was an intergroup association, $\chi^2(11, k = 571) = 17.72, p = .088$, and whether the target association was related to health or clinical issues, $\chi^2(11, k = 571) = 12.27, p = .343$.

The specific differences for the significant moderators are shown in Figure 4. Procedures that induce goals to weaken bias drove most of the moderator differences. These procedures produced stronger effect sizes in non-student samples ($g_{non-student} = -.44, g_{student} = -.24$), samples
with proportionally fewer White people ($g_{60\%\; \text{White}} = -.31$, $g_{100\%\; \text{White}} = -.07$), studies that used an IAT ($g_{\text{IAT}} = -.38$, $g_{\text{non-IAT}} = -.14$), studies with a pre-test implicit task ($g_{\text{pre-test}} = -.107$, $g_{\text{post-test\; only}} = -.23$), and studies that assessed a self-related association ($g_{\text{self}} = -.73$, $g_{\text{non-self}} = -.27$), though the 95% CI for this last difference overlapped slightly with 0. Future research could explore why such differences exist.

Student and non-student samples also tended to produce different effect sizes. In addition to the difference between student and non-student samples for studies using weaken goals procedures, student and non-student samples produced different effect sizes in studies that weakened associations indirectly ($g_{\text{non-student}} = -.08$, $g_{\text{student}} = -.28$) and that depleted cognitive resources ($g_{\text{non-student}} = -.15$, $g_{\text{student}} = -.32$). Finally, studies using an IAT produced stronger effects than non-IAT studies when they strengthened associations directly ($g_{\text{IAT}} = .25$, $g_{\text{non-IAT}} = .08$) and weakened associations indirectly ($g_{\text{IAT}} = -.28$, $g_{\text{non-IAT}} = -.12$), and studies that depleted a self-related association produced stronger effects than studies that did not ($g_{\text{self}} = .81$, $g_{\text{non-self}} = .16$).

**How do changes in implicit measures correspond with changes in explicit measures?**

To test whether the effects on implicit measures are consistent with effects on explicit measures, we fit a network meta-analytic model that allows the simultaneous analysis of two correlated outcomes (Achana et al., 2014; Efthimiou et al., 2015). This model revealed that effects on implicit measures differed from effects on explicit measures, $\chi^2(11, k = 570) = 30.58, p = .001$. Although effects on explicit measures were non-zero, $\chi^2(11, k = 570) = 68.03, p < .001$, they tended to be small by conventional standards ($g < .20$) and smaller than implicit effects. As shown in Figure 5, three of the eleven procedures had effects on explicit measures that were significantly smaller than their effects on implicit measures: weaken associations directly, $g = -.17$, 95% CI = [-.23, -.10], weaken associations indirectly, $g = -.13$, 95% CI = [-.21, -.05], and weaken goals, $g = -.11$, 95% CI = [-.21, -.03]. The rest of the procedures except for threat, affirmation, and negative affect had non-significantly smaller effects on explicit measures.

Explicit effect sizes tended to be less variable than implicit effect sizes, both in terms of the percentage of between-studies heterogeneity ($I^2_{\text{implicit}} = .797$, $I^2_{\text{explicit}} = .774$) and the effect size standard deviations ($\tau_{\text{implicit}} = .284$, $\tau_{\text{explicit}} = .238$).

To test whether implicit measure change mediated the effects of procedures on explicit measures and whether explicit measure change mediated the effects of procedures on implicit measures, we synthesized the indirect effects extracted from the correlation matrices from each study using two-stage meta-analytic structural equation modeling (Cheung & Chan, 2005; Cheung & Cheung, 2016). We modeled the differences between the indirect effects resulting from different procedure comparisons using a contrast-based approach, which represents direct comparisons using dummy codes and indirect comparisons using treatment contrasts (Salanti, Higgins, Ades, & Ioannidis, 2008). Because we only conducted these analyses with two-condition studies for which we knew the implicit effect size, explicit effect size, and the correlation between implicit and explicit measures, the results are based on fewer studies ($k =

---

9 One study was removed from this analysis because its within-studies variance-covariance matrix of effects on implicit and explicit measures was degenerate.
Figure 4. Moderation analyses. $k$ gives the number of studies that directly (or indirectly, listed in parentheses) compare the listed procedure and a neutral procedure for the displayed levels of the moderator. “Difference” represents the difference between the two moderator levels and its 95% CI. Higher effect sizes reflect greater increases in implicit measures compared to a neutral procedure. Where there was not enough data in one of the moderator levels for estimation, the overall model estimate is shown instead.
**Figure 5.** Forest plot of the consistency between effects on implicit and explicit measures. $g$ gives the implicit and explicit estimates; $g_I - g_E$ gives their difference. $k$ gives the number of studies with implicit and explicit measures that directly (or indirectly, listed in parentheses), compare the listed procedure and a neutral procedure. “$\chi^2$” gives the 1 df Wald $\chi^2$ test of the implicit-explicit difference, and “$p$” gives its $p$-value.

<table>
<thead>
<tr>
<th>Procedure category</th>
<th>$g$ [95% CI]</th>
<th>$g_I - g_E$</th>
<th>$k$</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weaken associations directly</td>
<td>-0.23 [-0.29, -0.17]</td>
<td>-0.17 [-0.23, -0.10]</td>
<td>0.07</td>
<td>55 (25)</td>
<td>3.3</td>
</tr>
<tr>
<td>Strengthen associations directly</td>
<td>0.21 [0.14, 0.28]</td>
<td>0.13 [0.05, 0.21]</td>
<td>0.07</td>
<td>37 (26)</td>
<td>1.4</td>
</tr>
<tr>
<td>Weaken associations indirectly</td>
<td>-0.23 [-0.29, -0.16]</td>
<td>-0.13 [-0.21, -0.05]</td>
<td>0.10</td>
<td>46 (14)</td>
<td>5.3</td>
</tr>
<tr>
<td>Strengthen associations indirectly</td>
<td>0.14 [0.05, 0.23]</td>
<td>0.05 [0.06, 0.16]</td>
<td>0.09</td>
<td>18 (12)</td>
<td>28.0</td>
</tr>
<tr>
<td>Goals to weaken bias</td>
<td>-0.28 [-0.36, -0.21]</td>
<td>-0.12 [-0.21, -0.03]</td>
<td>0.17</td>
<td>34 (7)</td>
<td>11.1</td>
</tr>
<tr>
<td>Goals to strengthen bias</td>
<td>0.15 [0.02, 0.27]</td>
<td>0.13 [-0.07, 0.33]</td>
<td>0.01</td>
<td>4 (7)</td>
<td>0.0</td>
</tr>
<tr>
<td>Threat</td>
<td>0.10 [0.00, 0.19]</td>
<td>0.04 [-0.08, 0.17]</td>
<td>0.05</td>
<td>23 (12)</td>
<td>0.7</td>
</tr>
<tr>
<td>Affirmation</td>
<td>-0.01 [-0.18, 0.16]</td>
<td>0.10 [-0.10, 0.29]</td>
<td>0.10</td>
<td>3 (10)</td>
<td>1.0</td>
</tr>
<tr>
<td>Positive affective state</td>
<td>-0.07 [-0.24, 0.09]</td>
<td>-0.02 [-0.18, 0.14]</td>
<td>-0.05</td>
<td>9 (4)</td>
<td>0.4</td>
</tr>
<tr>
<td>Negative affective state</td>
<td>-0.14 [-0.33, 0.04]</td>
<td>-0.23 [-0.50, 0.03]</td>
<td>0.09</td>
<td>2 (5)</td>
<td>0.5</td>
</tr>
<tr>
<td>Depletion</td>
<td>0.22 [0.07, 0.38]</td>
<td>0.02 [-0.25, 0.28]</td>
<td>0.21</td>
<td>5 (1)</td>
<td>2.1</td>
</tr>
</tbody>
</table>
187) than the full set of studies that contain an explicit task ($k = 260$). All values from this analysis can be interpreted as the product of a correlation and a semi-partial correlation.

As shown in Figure 6, the indirect effects are all quite small. A Wald $\chi^2$ test suggested that we could not reject the null hypothesis that the indirect effects of procedures on explicit measures through implicit measure change were zero, $\chi^2(10, k = 187) = 7.76, p = .735$. None of the individual estimates for the indirect effects were different from zero. These mediation results are not consistent with a causal relationship between change in implicit measures and change in explicit measure, although measurement and methodological issues in this meta-analysis could have obscured evidence for mediation (see General Discussion for elaboration). There was so little variation between studies in the magnitude of the indirect effects that the variation had to be fixed to zero for the models to converge. This last result suggests that it is highly unlikely that there are hidden moderators that would identify a subset of studies with evidence of a non-zero mediation effect.

Finally, we examined whether effect sizes were related to measurement correspondence between implicit and explicit tasks. Implicit and explicit effect sizes were related to measurement correspondence, $\chi^2(22, k = 258) = 39.61, p = .012$. Measurement correspondence did not explain the gap in effect sizes between implicit and explicit measures, $\chi^2(11, k = 258) = 11.73, p = .385$; less correspondent studies showed greater evidence for change than more correspondent studies for both implicit measures, $\chi^2(11, k = 258) = 25.38, p = .008$, and explicit measures, $\chi^2(11, k = 258) = 21.06, p = .033$. We attempted to fit a model testing whether the mediation effects in studies using higher correspondence implicit and explicit tasks were larger than those in studies with less correspondent tasks, but were unable to attain model convergence. We describe these analyses in more detail in our supplement at https://osf.io/awz2p/

**How do changes in implicit measures correspond with changes in behavior?**

We performed a similar set of analyses on behavior as we did on explicit measures. The procedures had a significant effect on behavior, $\chi^2(7, k = 487) = 23.42, p = .001$, though the size of these effects differed markedly from the implicit effects, $\chi^2(7, k = 487) = 23.75, p = .001$. As shown in Figure 7, the six procedures that invoked threat produced a small-to-moderate overall effect on behavior that may have driven the overall effect, $g = .39, 95\%$ CI = [.14, .64]. These six procedures did not have an overall effect on implicit measures, $g = .05, 95\%$ CI = [-.06, .16]. The only other procedure category with a significant effect was weaken associations directly, $g = -.10, 95\%$ CI = [-.20, -.01], which had an “trivial” effect size by conventional standards (Hyde, 2005, 2014). All other procedures produced behavioral effects that were smaller than their corresponding effects on implicit measures. Behavioral effects were less variable than implicit effects, both measured in terms of the percentage of between-studies heterogeneity ($I^2_{\text{implicit}} = .787, I^2_{\text{behavior}} = .692$) and the effect size standard deviations ($\tau_{\text{implicit}} = .302, \tau_{\text{behavior}} = .269$).

---

10 Studies with affirmation, positive or negative affect, or depletion procedures were excluded from this analysis because there were no studies with behavioral tasks that used these procedures. An additional study was removed from this analysis because its within-studies variance-covariance matrix of effects on implicit and behavioral bias was degenerate.
Figure 6. Indirect effects (in the conventional mediation framework, the effect $ab$) of procedures on explicit measures through changes in implicit measures. $k$ gives the number of studies that directly (or indirectly, listed in parentheses) compare the listed procedure and a neutral procedure.

<table>
<thead>
<tr>
<th>Procedure category</th>
<th>$k$</th>
<th>$ab$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weaken associations directly</td>
<td>33 (24)</td>
<td>-0.002 [-0.006, 0.002]</td>
</tr>
<tr>
<td>Strengthen associations directly</td>
<td>23 (24)</td>
<td>0.001 [-0.003, 0.005]</td>
</tr>
<tr>
<td>Weaken associations indirectly</td>
<td>22 (14)</td>
<td>-0.002 [-0.008, 0.003]</td>
</tr>
<tr>
<td>Strengthen associations indirectly</td>
<td>6 (12)</td>
<td>0.001 [-0.005, 0.007]</td>
</tr>
<tr>
<td>Goals to weaken bias</td>
<td>15 (6)</td>
<td>-0.007 [-0.018, 0.003]</td>
</tr>
<tr>
<td>Goals to strengthen bias</td>
<td>1 (7)</td>
<td>-0.003 [-0.016, 0.011]</td>
</tr>
<tr>
<td>Threat</td>
<td>19 (11)</td>
<td>0.001 [-0.006, 0.008]</td>
</tr>
<tr>
<td>Affirmation</td>
<td>0 (9)</td>
<td>0.004 [-0.009, 0.017]</td>
</tr>
<tr>
<td>Positive affective state</td>
<td>5 (4)</td>
<td>-0.012 [-0.025, 0.000]</td>
</tr>
<tr>
<td>Negative affective state</td>
<td>0 (4)</td>
<td>-0.013 [-0.026, 0.001]</td>
</tr>
<tr>
<td>Depletion</td>
<td>5 (1)</td>
<td>-0.006 [-0.020, 0.008]</td>
</tr>
</tbody>
</table>
Figure 7. Forest plot of the consistency between effects on implicit and behavioral measures. \( g \) gives the implicit and behavioral estimates; \( g_I - g_B \) gives their difference. \( k \) gives the number of studies with implicit and behavioral measures that directly (or indirectly, listed in parentheses) compare the listed procedure and a neutral procedure. “\( \chi^2 \)” gives the 1 df Wald \( \chi^2 \) test of the implicit-behavioral difference, and “\( p \)” gives its \( p \)-value.

<table>
<thead>
<tr>
<th>Procedure category</th>
<th>( g [95% \text{ CI}] )</th>
<th>( g_I - g_B )</th>
<th>( k )</th>
<th>( \chi^2 )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>weaken associations directly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>-0.23 [-0.30, -0.16]</td>
<td>-0.13</td>
<td>33 (14)</td>
<td>7.2</td>
<td>0.01 *</td>
</tr>
<tr>
<td>behavior</td>
<td>-0.10 [-0.20, -0.01]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strengthen associations directly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>0.21 [0.14, 0.29]</td>
<td>0.11</td>
<td>22 (11)</td>
<td>4.2</td>
<td>0.04 *</td>
</tr>
<tr>
<td>behavior</td>
<td>0.10 [-0.01, 0.21]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weaken associations indirectly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>-0.23 [-0.30, -0.16]</td>
<td>-0.11</td>
<td>9 (16)</td>
<td>2.3</td>
<td>0.13</td>
</tr>
<tr>
<td>behavior</td>
<td>-0.12 [-0.27, 0.02]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strengthen associations indirectly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>0.13 [0.03, 0.23]</td>
<td>0.17</td>
<td>2 (11)</td>
<td>4.1</td>
<td>0.04 *</td>
</tr>
<tr>
<td>behavior</td>
<td>-0.04 [-0.22, 0.14]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>goals to weaken bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>-0.29 [-0.37, -0.21]</td>
<td>-0.17</td>
<td>6 (4)</td>
<td>2.9</td>
<td>0.09</td>
</tr>
<tr>
<td>behavior</td>
<td>-0.12 [-0.32, 0.07]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>goals to strengthen bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>0.14 [0.01, 0.28]</td>
<td>0.16</td>
<td>2 (3)</td>
<td>1.5</td>
<td>0.22</td>
</tr>
<tr>
<td>behavior</td>
<td>-0.02 [-0.29, 0.24]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>threat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit</td>
<td>0.06 [0.06, 0.16]</td>
<td>-0.34</td>
<td>6 (0)</td>
<td>7.1</td>
<td>0.01 *</td>
</tr>
<tr>
<td>behavior</td>
<td>0.38 [0.14, 0.64]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As shown in Figure 8, we estimated whether implicit measure change mediated the effects of procedures on behaviors. As with explicit measures, this analysis is based on a set of samples (\( k = 63 \)) that is smaller than the set of samples that contain a behavioral task (\( k = 94 \)) because it only includes two-condition studies that had complete data. In the aggregate, procedures did not produce significant indirect effects, \( \chi^2(7, k = 63) = 5.19, p = .637 \). Follow-up examination of the individual indirect effects revealed that none were significantly non-zero. These mediation results are not consistent with a causal relationship between change in implicit measures and change in behavior, although measurement and methodological issues in this meta-analysis could have obscured evidence for mediation (see General Discussion). As with the indirect effects on explicit measures, there was so little variation between studies in the size of the indirect effects that the variation had to be fixed to zero for the models to converge, once again suggesting that there are no hidden moderators that would identify a subset of studies with stronger evidence of a non-zero mediation effect.
Figure 8. Indirect effects (in the conventional mediation framework, the effect \( ab \)) of procedures on behavioral measures through changes in implicit measures. \( k \) gives the number of studies that directly (or indirectly, listed in parentheses) compare the listed procedure and a neutral procedure.

<table>
<thead>
<tr>
<th>Procedure category</th>
<th>( k )</th>
<th>( ab ) [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weaken associations directly</td>
<td>19 (9)</td>
<td>0.000 [-0.005, 0.004]</td>
</tr>
<tr>
<td>Strengthen associations directly</td>
<td>9 (9)</td>
<td>0.002 [-0.005, 0.009]</td>
</tr>
<tr>
<td>Weaken associations indirectly</td>
<td>8 (10)</td>
<td>-0.002 [-0.010, 0.006]</td>
</tr>
<tr>
<td>Strengthen associations indirectly</td>
<td>0 (10)</td>
<td>-0.002 [-0.012, 0.007]</td>
</tr>
<tr>
<td>Goals to weaken bias</td>
<td>2 (3)</td>
<td>-0.052 [-0.118, 0.013]</td>
</tr>
<tr>
<td>Goals to strengthen bias</td>
<td>0 (3)</td>
<td>-0.033 [-0.104, 0.039]</td>
</tr>
<tr>
<td>Threat</td>
<td>3 (0)</td>
<td>0.003 [-0.013, 0.018]</td>
</tr>
</tbody>
</table>

We also tested whether effect sizes were related to measurement correspondence, whether the behavior was deliberate, and whether the behavior was impulsive. Past meta-analyses of implicit measures have remarked on how different subjective coding methods on variables like these could lead to dramatically different conclusions (Cameron et al., 2012; Oswald et al., 2013). We encountered similar issues, as most studies did not report on the information necessary to make an objective determination. As such, these results should be interpreted with caution.

We found that implicit and behavioral effect sizes were not related to measurement correspondence, \( \chi^2(10, k = 92) = 13.59, p = .193 \), or deliberateness, \( \chi^2(10, k = 90) = 11.49, p = .321 \). However, effect sizes were related to impulsiveness, \( \chi^2(10, k = 90) = 18.38, p = .049 \), but with weak evidence barely below the .05 significance criterion (Benjamin et al., 2018). We next examined whether correspondence, impulsiveness, or deliberateness explained the difference in effect sizes between implicit and behavioral measures and found that they did not, correspondence \( \chi^2(5, k = 92) = 10.59, p = .060 \), impulsiveness \( \chi^2(5, k = 90) = 5.90, p = .316 \), deliberateness \( \chi^2(5, k = 90) = 1.57, p = .904 \). Compared to studies with non-impulsive behaviors, studies with impulsive behaviors showed greater evidence for change on their behavior, \( \chi^2(5, k = 90) = 16.60, p = .005 \), but not their implicit measures, \( \chi^2(5, k = 90) = 5.17, p = .396 \). We attempted to fit models testing whether these three variables were associated with the size of the mediation effects but were unable to fit a model that converged. We describe these analyses in more detail at [https://osf.io/awz2p/](https://osf.io/awz2p/).

Is there evidence that the size of reported effects is biased?
We tested for biases in effect sizes by assessing funnel plot asymmetry (Egger et al., 1997), estimating weight-function models (Vevea & Hedges, 1995), conducting trim-and-fill (Duval & Tweedie, 2000), and by assessing whether effect sizes varied by publication status, year, or geographic location.11

Funnel plots show study effect sizes plotted against their standard errors (Egger et al., 1997). Funnel plots of an unbiased literature have a fan shape, with studies centering around a single effect size, regardless of precision, but with a greater scatter around the effect size in low-precision studies. Bias causes asymmetry in funnel plots by preventing a subset of low-precision studies (e.g., those with non-significant results) from entering the meta-analysis. Comparison-adjusted funnel plots are funnel plots adapted to network meta-analysis (Chaimani et al., 2013). Although they cannot accommodate multiple effects from the same study, they can accommodate studies that examine different sets of comparisons between procedures. They account for these different comparisons by subtracting the relevant meta-analytic comparison estimate (e.g., threat vs. neutral, weaken goals vs. neutral) from each study estimate prior to plotting. As in a normal funnel plot, one can then examine the comparison-adjusted plots for asymmetry, which suggests that some process differentially affected high and low precision studies (e.g., publication bias).

To select a set of two-group studies (published and unpublished) in which most researchers would make similar predictions, we made the following three generic predictions. First, the weaken associations directly, weaken associations indirectly, and weaken goals procedures will reduce response bias on implicit, explicit, and behavioral measures relative to a neutral procedure. Second, the strengthen associations directly, strengthen associations indirectly, strengthen goals, and deplete resources procedures will increase response bias relative to a neutral procedure. Third, procedures in the first group will result in less response bias than procedures in the second.

The funnel plots of the comparison-adjusted effect sizes for these studies on implicit, explicit, and behavioral measures are shown in Figure 9. The figure reveals asymmetry in all plots in that high-precision effect sizes tended to be smaller than their corresponding overall meta-analytic estimates. This observation was supported by the results of mixed-effect regression analyses (Sterne & Egger, 2005) testing the relationship between implicit standard errors and effect sizes, $z = 3.60$, $p < .001$ and explicit standard errors and effect sizes, $z = 2.84$, $p = .005$. There was no significant relationship between the behavioral standard errors and effect sizes, $z = 1.29$, $p = .196$. However, the relationship between standard errors and behavioral effect sizes was estimated with much less precision than the implicit and explicit relationships. If the funnel plot asymmetry is caused by processes that systematically prevent small, non-significant effect sizes from entering the meta-analysis (e.g., publication bias, $p$-hacking), this suggests that implicit and explicit effects in this meta-analysis are inflated relative to their population values.

---

11 We considered implementing other bias detection methods, such as $p$-curve analysis (Simonsohn, Nelson, & Simmons, 2014), but ultimately did not because they depend on the assumption of homogeneity and have not yet been adapted to examining bias in a network of interventions where heterogeneity is expected a priori (for a review, see Efthimiou et al., 2016).
**Figure 9.** Comparison-adjusted funnel plots of effect sizes vs standard errors for implicit, explicit, and behavioral measures. Positive numbers are more extreme relative to the meta-analytic comparison a study contributes to and negative numbers less extreme. The red line represents the fit from a mixed-effects regression; a line that departs from the vertical suggests the presence of small-study bias.

We also examined bias in effect sizes with weight function models and trim-and-fill. We fit weight function models (Vevea & Hedges, 1995) using the `weightr` package (Coburn & Vevea, 2017) to test whether studies with p-values greater than .05 occurred less frequently than one would expect based on sampling error, adding moderators for the comparison tested by each study to account for the extra heterogeneity due to the fact that different studies were testing different procedures. The results are partially consistent with those of the comparison-adjusted funnel plots: implicit effects with computed p-values greater than .05 were .37 times less likely to occur than one would expect based on sampling error, 95% CI = [.23, .52], whereas behavioral effects with p-values greater than .05 were not significantly different from p-values less than .05, b = .57, 95% CI = [.00, 1.20]. Unlike the funnel plot analyses, explicit effects with p-values greater than .05 did not occur at significantly different rates than p-values less than .05, b = 2.79, 95% CI = [.89, 4.70].

We also used the trim-and-fill method (Duval & Tweedie, 2000), which suggested that 56 studies were missing from our set of implicit studies, but that no explicit or behavior studies were missing. These last results should be interpreted with extreme caution as simulation evidence suggests that trim-and-fill is inadequate at detecting and correcting for small-study effects (Rücker, Carpenter, & Schwarzer, 2011).

Funnel plot analyses, weight function models, and trim-and-fill do not distinguish between the many processes that could lead to bias in effect sizes. Potential causes are better distinguished with moderator analyses. We conducted moderator analyses using publication year to test for decline effects (Jennions & Møller, 2002), publication status to test for publication bias (Stern & Simes, 1997), and geographic region to test for United States bias (Fanelli & Ioannidis, 2013).

---

12 These coefficients are multiplicative, and therefore significant if their 95% CI does not include 1.
Implicit effect sizes varied by publication year, $\chi^2(11, k = 463) = 25.51, p = .008$. As shown in Figure 10, there was a general tendency for more recent studies to yield (non-significantly) smaller effect sizes. There were two exceptions: strengthen associations indirectly, for which effect sizes remained constant across all publication years, $b = .006$, 95% CI = [-.025, .038], and goals to weaken bias, for which there was a growth effect rather than a decline effect – more recent studies have larger (more negative) effect sizes, $b = -.030$, 95% CI = [-.052, -.008]. This last relationship may be driven by research showing that response biases on implicit tasks are sensitive to strategic responding (e.g., implementation intentions to reduce bias on a shooter bias task, Mendoza, Gollwitzer, & Amodio, 2010, instructions to Germans to fake a pro-Turkish IAT score, Fiedler & Bluemke, 2005). Early studies suggested that implicit measures were resistant to strategic responding (Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim, 2003), whereas more recent studies have suggested that strategic responding is possible, particularly with sufficiently specific instructions (Fiedler & Bluemke, 2005; Lai et al., 2014; 2016; Stewart & Payne, 2008). Contrary to evidence from other areas of research (Stern & Simes, 1997; Fanelli & Ioannidis, 2013), implicit effect sizes did not depend on publication status, $\chi^2(11, k = 571) = 17.93, p = .083$, or geographic location, $\chi^2(11, k = 571) = 6.09, p = .867$. Are the results robust to an alternative coding scheme?
The main procedure coding scheme did not distinguish between procedures that present new information (learning) from procedures that re-activate old information that is already in memory (context). For example, learning about the statistical link between cigarette smoking and cancer (Smith & De Houwer, 2015) may have entirely different implications for psychological change than the context-based influence of smelling cigarettes in the air (Glock, Kovacs, & Unz, 2014). Basic research on the distinction between change in context-free general representations and change in contextualized representations suggest that this distinction has implications for the duration and generalizability of psychological change (Gawronski & Cesario, 2013; Gawronski et al., 2010; 2015). To understand whether this distinction is relevant for the current results, we split the four procedure categories that attempted to directly or indirectly change associations into eight categories that distinguished between the presentation of new and already-known information. As almost no papers explicitly tested the difference between procedures that evoke learning vs context, the information necessary to make this distinction clearly was seldom described in the paper. For this reason, although we were able to make this distinction with an acceptable level of reliability (κ = .71), making this distinction in theoretically valid way may be impossible short of conducting new experiments explicitly designed to examine this distinction.

Nevertheless, we tested the robustness of our results to the distinction between learning and context by re-fitting our primary statistical models and testing whether the procedures involving learning produced different effect sizes than the procedures involving context (see the supplement at https://osf.io/awz2p/ for details about specific models). Out of 19 statistical models, we found that the learning and context effects differed in only three cases: implicit moderation analyses involving student vs. non-student samples, post-only designs vs. pre-post designs, and behavioral moderation analyses examining whether the measure was deliberate or non-deliberate. The patterns in each of these models were not consistent or easily interpretable, suggesting false-positive results or hidden variables. These findings suggest that the main results are robust to this alternative coding scheme.

**General Discussion**

Our meta-analysis is the first large-scale quantitative synthesis of research on change in implicit measures. We found that implicit measures can be changed across many areas of study, populations, implicit tasks, and research designs. The type of approach used to change implicit measures mattered greatly. Some procedures were effective at changing implicit measures, whereas others were not. Procedures to change implicit measures produced smaller changes in explicit measures and behavior, and we found no evidence that changes in implicit measures mediate changes in explicit measures and behavior.

**Relative effectiveness of procedures to change implicit measures**

We developed a taxonomy for understanding how procedures to change implicit measures differed. Using this taxonomy, we found that procedures that directly or indirectly targeted associations, depleted mental resources, or induced goals all changed implicit measures relative to neutral procedures. In contrast, procedures that induced threat, affirmation, or affective states had small and/or inconsistent effects. These results support the theoretical portrayal of automatically retrieved associations as sensitive to pairings of information in the
social environment (Gawronski & Bodenhausen, 2006). These results also support the importance of goal-directed motivation and cognitive resources in changing the expression of automatically retrieved associations (Fazio & Olson, 2014; Gawronski & Payne, 2010; Devine, 1989).

The procedures that produced robust effects on implicit measures had average effects that were relatively small by conventional standards (Hyde, 2005) and below the median effect size in social psychology (Richard, Bond, & Stokes-Zoota, 2003). All three of the tests we conducted to examine bias in the implicit effects suggested that the population effects of these procedures may be even smaller than our meta-analytic estimates due to publication bias, p-hacking, and/or other processes.

**Generalizability of implicit measure change**

We also uncovered evidence of large variation in the size of the effects produced by procedures to change implicit measures. Some of the sources of this variation reveal complexities in evaluating the impact of the procedures on implicit measures. First, researchers’ choices of samples have constrained the generalizability of the available evidence (Henrich, Heine, & Norenzayan, 2010). Most studies have been conducted with samples whose demographic characteristics (students, mostly White, mostly female) strongly resemble those of Introductory Psychology classrooms in the United States. Although the gender composition of the sample was not associated with the size of effects, both the racial composition of the samples and whether the samples were drawn from university student populations were. Student samples in particular produced different effect sizes than non-student samples for three of the nine procedure comparisons that we examined (strengthen associations directly vs. neutral, weaken associations indirectly vs. neutral, goals to weaken bias vs. neutral).

Because studies with university student samples often address different research questions than studies with non-university student samples and because university students are psychologically different from the general population (Henrich et al., 2010; Sears, 1986), the precise cause of these different effect sizes is unclear. Regardless, these results suggest that it would be prudent to directly test whether the effects of manipulations are generalizable to other populations. Combating societal problems such as discrimination and addiction requires exploration of how the problems operate outside of the college campus, and answering questions of human nature depends on sampling from a population that represents humankind.

Another limit to generalizability is a lack of research interest in change beyond the confines of a single experimental session. The present meta-analysis speaks more to the processes that change implicit measures in the short-term rather than to processes that change implicit measures in the long-term. Only 17 (3.0%) samples used procedures that took longer than one session to complete. Only 38 (6.6%) samples in the meta-analysis collected longitudinal outcomes and therefore had the opportunity to examine whether the procedures they investigated produce long-term changes. Short-term changes in implicit measures do not necessarily generalize to longer-term changes (Devine, Forscher, Austin, & Cox, 2012; Forscher et al., 2017; Forscher & Devine, 2014; Lai et al., 2016; Lai, Hoffman, & Nosek, 2013; Miller, Dannals & Zlatev, 2017). This issue is of critical importance given theorizing that automatically retrieved
associations are created and sustained by repeated pairings of information in the social environment. That means that without active efforts to sustain short-term shifts created in the lab, these shifts are likely to be wiped away upon re-exposure to the social environment (Forscher et al., 2017; cf. De Houwer, 2009; Mann & Ferguson, 2017). In fact, one recent series of studies found that nine interventions that reduced response biases on implicit tasks immediately showed little to no lasting impact days later (Lai et al., 2016). What processes determine whether a shift in implicit measures will be temporary or long-lasting? When will a shift in implicit measures translate into a more permanent change? Theory and practice-oriented researchers alike must contend with these questions.

Effect sizes also differed according to a study’s methodological features. Studies using an IAT produced effects that were often larger than studies that did not, and studies with a pretest post-test design that induced a goal to weaken bias produced larger effects than studies that only included a post test assessment. The large IAT effects could be driven by the IAT’s reliability, which is typically higher than the reliability of most other implicit tasks (Bar-Anan & Nosek, 2014; Bosson et al., 2000).

The effects of interventions did not vary much based on their topic. Studies that targeted evaluative associations did not differ from studies that targeted conceptual associations, and effect sizes did not differ as a function of domain (e.g., intergroup relations, clinical/health).

**Implicit measures and explicit measures**

Most studies of the relationship between the implicit and explicit measures are observational studies that administer implicit and explicit tasks within the same session. These relationships can be very low or very high, and are highest – when using the IAT at least – when people’s thoughts about the concepts are well-elaborated, when the explicit measure is more affective, when the topic of study is political preferences, when the concepts are diametrically opposed (e.g., liberals vs. conservatives), and when people perceive that their opinions about the concepts are distinct from the opinions of others (Cameron et al., 2012; Greenwald et al., 2009; Hofmann et al., 2005; Nosek, 2005). Although it was not the primary purpose of our meta-analysis, we found that the correlation between implicit and explicit measures in our sample of experimental studies was low ($r_{I,E} = .14$). This is a marked difference from the median ($r_{I,E} = .38$) of large-sample studies ($N > 100,000$) investigating highly heterogeneous topics in highly heterogeneous samples. In fact, compared to 95 examined topics, the estimate from this meta-analysis was smaller than all but one (Nosek & Hansen, 2008).

There are good reasons expect a different correlation in experimental studies than in observational studies, as experimental manipulations could influence the correlation between implicit and explicit measures. For example, manipulations could affect levels of systematic or random measurement error or change the rank ordering of performance in one outcome but not the other outcome.

The available studies also tended to focus on a limited range of topics and samples. For example, the most common topic in this meta-analysis was intergroup relations (63.4% of studies), an area known for low implicit-explicit correlations in observational studies (Hofmann
et al., 2005; Nosek, 2005, 2007). This topical bias is understandable considering that most research applications for changing implicit measures is for topics that elicit implicit responses that are unwanted or distinct from deliberately reported explicit evaluations. Many samples were also composed of predominantly White university students. This homogeneous sampling may have constrained the magnitude of the correlation between implicit and explicit measures beyond what might be expected due to the causal impact of experimental manipulations.

Our focus on randomized studies gave us an opportunity to go beyond correlational evidence by examining whether procedures that attempt to change implicit measures also produce change in explicit measures. We found that many of the procedures that change implicit measures also produce change in explicit measures, though the magnitude of change in explicit measures was weaker and less variable. Simultaneously, there was no evidence that changes in implicit and explicit measures were mediated by each other. One possibility suggested by these data is that there is no relationship between changes in implicit and explicit measures. This possibility would reduce support for theoretical perspectives that posit interdependence between automatic and deliberate processes that are presumed to underlie implicit and explicit measures (e.g., Gawronski & Bodenhausen, 2006; c.f. Smith & DeCoster, 2000). However, even if this is true, we cannot eliminate the possibility that the relationship is stronger in other samples or topics.

It is not possible from these data to determine whether increasing diversity in samples, designs, and topics would yield substantively different mediation results. The most productive next step is to evaluate these possibilities directly. There are some hints that such investigations would yield stronger mediation evidence. For example, Smith, Ratliff, and Nosek (2012) had large samples of participants (N’s = 732; 621) form attitudes toward novel policy proposals that were randomly attributed to Democrats or Republicans. Implicit and explicit attitudes toward the plans were strongly correlated (r’s = .48, .51/.59) and implicit attitudes fully mediated the effect of the experimental intervention on explicit attitudes, but not the reverse, both immediately and 5 days after the intervention.

This example was not included in this meta-analysis because we only examined studies of pre-existing associations. As a consequence, this and all other studies of the formation of new associations were excluded. This creates an interesting mystery to be solved. The association formation literature provides substantial experimental evidence for the interdependence of automatically and deliberately retrieved associations (e.g., Gawronski & Bodenhausen, 2006, 2011; Gawronski & LeBel, 2008; Gawronski, Rydell, Vervliet, & De Houwer, 2010; Moran, Bar-Anan, & Nosek, 2015; Ranganath & Nosek, 2008). In contrast, this meta-analysis on pre-existing associations provides little evidence of interdependence. Whatever the explanation, resolving the apparent discrepancy between research on new and pre-existing associations provides an exciting opportunity to advance theory about implicit social cognition.

**Implicit measures and behavior**

Previous investigations of implicit-behavior relations have also relied on observational studies. Meta-analytic estimates of this relationship vary substantially (Greenwald et al., 2009 $r_{IB} = .27$; Cameron et al., 2012 $r_{IB} = .28$; Kurdi et al., 2018 $r_{IB} = .10$; Oswald et al., 2013 $r_{IB}$...
The correlations between implicit measures and behavior tend to be smallest for topics in which automatic and deliberate processes are least likely to facilitate each other, such as race relations (Greenwald et al., 2009; Kurdi et al., 2018). The overall correlation between implicit measures and behavior in our meta-analysis was small and closer to the estimates in the meta-analyses on these topics ($r_{IB} = .09$).

On the surface, this research is about prediction, but of course, the interest is also about causation. Indeed, many researchers use evidence of correlations between implicit measures and behavior to argue for the causal importance of automatically retrieved associations (e.g., Banaji, Bhaskar, & Brownstein, 2015; Devine et al., 2012; Dovidio, Kawakami, & Gaertner, 2002; Green et al., 2007; Kang & Banaji, 2006). For example, Devine, Forscher, Austin, and Cox (2012, p. 1267) argue on the basis of correlational studies that “accumulating evidence reveals that implicit biases are linked to discriminatory outcomes ranging from the seemingly mundane, such as poorer quality interactions (McConnell & Leibold, 2001), to the undeniably consequential, such as constrained employment opportunities (Bertrand & Mullainathan, 2004) and a decreased likelihood of receiving life-saving emergency medical treatments (Green et al., 2007). [...] [Implicit bias] leads people to be unwittingly complicit in the perpetuation of discrimination.”

Of course, correlations between variables can be produced by many relationships besides ones that are causal. To get closer to questions of causality, we looked at whether changes in implicit measures correspond with and mediate changes in behavior in our sample of randomized experiments. We found that the effect of procedures on behavior were trivial by conventional standards, with the exception of threat which had a small-to-moderate effect on behavior. We found no evidence that changes in implicit measures mediate changes in behavior.

The lack of evidence for mediation is difficult to reconcile with the correlational evidence. One limit to generalizability is the relatively small number of studies examining change in behavior ($k = 63$) with usable information for mediation analysis. Other limits include the heavy reliance on White student samples, single-session manipulations, and a narrow range of topics. Nevertheless, the lack of an observed effect is a clarion call that demands more direct, high-powered investigation of relations between changing implicit measures and behavior. Even if the relationship between changes in implicit measures and changes in behavior is truly larger in domains, samples, and manipulations that were not included in this meta-analysis, our results suggest some constraints on the conditions under which changing implicit measures will predict or cause corresponding changes in behavior.

Potential explanations for implicit measures’ relationships with explicit measures and behavior

Even if we accept that our explanations of our findings regarding the explicit and behavioral measures do not generalize to all samples and topics, we are left with specifying what those explanations are. We offer four possibilities.

First, our inclusion criteria for explicit and behavioral tasks may have led to the inclusion of measures that should not be theoretically expected to change after a change in automatically
retrieved associations. We included explicit and behavioral tasks that appeared to assess the same associations as the study’s implicit task, regardless of whether performance on that task was expected to change after the manipulation. For example, if the implicit task was a Black/White good/bad IAT, we included any explicit or behavioral task that connected race and valence. Eligible explicit tasks ranged from a simple feeling thermometer that assesses perceived warmth toward Whites vs. Blacks (Rudman, Dohn, & Fairchild, 2007) to the Symbolic Racism Scale that assesses the degree to which participants blame Black people for their current social standing (Inzlicht, Gutsell, & Legault, 2012). Eligible behavioral tasks ranged from how close a person sits to a Black confederate (Mann & Kawakami, 2012) to decisions about donating to children in South African vs. Colombian slums (Schwab & Greitemeyer, 2015). If the conditions under which change in automatically retrieved associations influence deliberately retrieved associations and behavior are narrow, our inclusion criteria may not have been sensitive to these narrow conditions.

To address this concern, we examined potential moderators of the relationship between implicit measures and explicit/behavioral measures and found mostly null effects. However, these between-study moderator analyses were limited by the procedural information reported in methods sections, which constrains what theoretical distinctions could be made during coding. Addressing this will require primary studies designed to examine specific theoretical distinctions. These moderator analyses were also limited by procedural differences between studies that could reduce power to detect effects due to between-studies error variance. Addressing this will require primary studies or meta-analyses of studies that were specifically designed to examine the relevant theoretical distinctions (e.g., Cameron et al., 2012).

Second, perhaps confounds introduced after the manipulations obscured the evidence for mediation. Statistical mediation analysis relies on the untestable assumption of a lack of confounding of the post-manipulation mediator-outcome relationship (Bullock, Green, & Ha, 2010). Most, but not all, sources of confounding will overstate the evidence for mediation (Bullock et al., 2010). However, confounding that reduces evidence for mediation could explain the null results. That may happen, for example if a second mediator that opposes the causal influence of automatically retrieved associations was also changed by many of the procedures examined in the meta-analysis. We cannot rule out this explanation, but we also cannot identify what these confounds would be.

Third, measurement issues may obscure the evidence for mediation within our studies. Almost all psychological tasks assess latent constructs indirectly through behavioral responses (Borsboom, 2006), and implicit tasks are no exception (Calanchini & Sherman, 2013; Conrey et al., 2005; Payne, 2001). Performance on implicit tasks is affected by an amalgam of processes, including associative processes, measurement error, and non-associative processes, such as task-switching ability, recoding, inhibition of impulses, and guessing (Calanchini et al., 2013; 2014; Klauer & Mierke, 2005). High levels of measurement error, as is characteristic of implicit tasks (Bosson et al., 2000; Buhrmester, Blanton, & Swann, 2011; Olson & Fazio, 2002) could obscure
Meta-analysis of change in implicit measures

Evidence that changes in automatically retrieved associations mediate changes in other processes.13

It is also possible that many of the procedures we examined produced change in implicit measures through non-associative processes. At least some of the procedures did. For example, a subset of studies that used goals to strengthen or weaken bias gave participants instructions to strategically respond or fake on an implicit task (e.g., Banse, Seise, & Zerbes, 2001; Fiedler & Bluemke, 2005). If many of our procedures produced change through non-associative processes, our analyses would bear on the effectiveness of these non-associative processes for changing explicit measures and behavior rather than the effectiveness of automatically retrieved associations. Without tools that isolate the contributions of associative and non-associative processes, we cannot definitively rule this possibility out.

Fourth, perhaps automatically retrieved associations really are causally inert. Accepting this conclusion would force reevaluation of some of the central assumptions that drive research on implicit social cognition. One such attempt in the intergroup domain is the “bias of crowds” model (Payne et al., 2017), which interprets mental associations as primarily a function of situational factors that somehow “add up” across people and time to exert a causal force on behavior. We entertain an even stronger proposal: instead of acting as a “cognitive monster” that inevitably leads to bias-consistent thought and behavior (e.g., Bargh, 1999; Tajfel, 1982), automatically retrieved associations reflect the residual “scar” of concepts that are frequently paired together within the social environment and do not have much causal force on their own. Similar to the bias of crowds model, automatically retrieved associations in the scar interpretation are a side effect of living in a particular social environment. In contrast to the bias of crowds model, the scar interpretation suggests that changes in automatically retrieved associations are epiphenomenal rather than changes in the mental processes that drive either deliberately retrieved associations or behavior.

This is not to say that the implicit measurement would be unproductive even under the scar interpretation. Demographic variables such as life expectancy are often used to predict other consequential outcomes within a population, despite lacking causal force themselves. By the same token, implicit measures could be used to predict the prevalence of certain judgments or behaviors within a population. However, under this interpretation, though the presence of an response biases on implicit tasks would speak to the structure of the social environment, efforts to change behavior by changing implicit measures would be misguided. It would be more effective to rid the social environment of the features that cause biases on behavioral and cognitive outcomes (Beaman, Duflo, Pande, & Topalova, 2012) or equip people with strategies to resist the environment’s biasing influence (Cohen & Sherman, 2014; Devine et al., 2012) rather than trying to alter the response biases themselves.

13 Measurement error in implicit tasks would not explain the trivially sized effects of procedures on behavioral outcomes, although measurement error in behavioral tasks might. Recent meta-analyses (Carlsson & Agerström, 2016; Kurdi et al., 2018) found that many behavioral tasks in correlational research on the IAT and discrimination lacked validity and reliability. Many of the behavioral tasks in this meta-analysis appeared to suffer from similar measurement issues. For example, many behavioral outcomes were based on as a single behavior (rather than an aggregate of multiple behaviors) and were not based on standardized procedures where the validity and reliability is well-known.
Presently, the scar interpretation is an incomplete account of the existing evidence on implicit social cognition. Although the scar interpretation of automatically retrieved associations explains correlations between implicit measures, explicit measures, and behavior as resulting from the shared cause of the social environment, this interpretation is nonspecific and does not explain why certain correlations between implicit measures and other variables are stronger than others. For example, well-elaborated concepts have stronger levels of convergence between implicit and explicit measures (Nosek, 2005), and people who have higher levels of working memory have lower levels of convergence between implicit measures and behavior (Friese, & Schmitt, 2008; Hofmann, Gschwendner, Friese, Wiers, & Schmitt, 2008; for a review, see Perugini, Richetin, & Zogmaister, 2010). A non-causality account would also have to integrate studies on novel associations which, at least in the case of explicit measures, provide stronger evidence for mediation (e.g., Gawronski & Bodenhausen, 2006, 2011; Gawronski & LeBel, 2008; Gawronski et al., 2010; Moran et al., 2015; Ranganath & Nosek, 2008).

The present meta-analysis is insufficient to distinguish between the competing explanations for our findings. Distinguishing between these explanations requires new evidence, possibly using a new paradigm. Ideally, this paradigm would involve a procedure that produces a robust and unambiguous causal impact on the automatically retrieved associations that underlie implicit measures, ideally in multiple domains. If this paradigm also creates changes in deliberatively retrieved associations and behavior that are themselves associated with the changes in automatically retrieved associations, this will provide supportive, though not definitive, evidence as to the downstream impacts of changing automatically retrieved associations (Bullock, Green, & Ha, 2010). To find such a paradigm, researchers might start with domains, such as political behavior, in which implicit, explicit, and behavioral measures are more intercorrelated (e.g., Ajzen & Fishbein, 2010; Greenwald et al., 2009; Hofmann et al., 2005; Kurdi et al., 2018; Nosek, 2005; 2007) as opposed to domains in which those relations are comparatively weak. Doing so would enable high-powered investigations of the impact of change interventions and mediating relationships among implicit, explicit, and behavioral measures (Smith et al., 2012). This would provide a first step toward resolving the theoretical and empirical puzzles raised by the present research.

**Conclusion**

This meta-analysis found that implicit measures can be changed and identified the approaches that are most successful in doing so. However, we found little evidence that changes in implicit measures translated into changes in explicit measures and behavior, and we observed limitations in the evidence base for implicit malleability and change.

These results produce a challenge for practitioners who seek to address problems that are presumed to be caused by automatically retrieved associations, as there was little evidence showing that change in implicit measures will result in changes for explicit measures or behavior. This is particularly true for the domains of greatest interest to many practitioners – intergroup bias, health psychology, and clinical psychology. Our results suggest that current interventions that attempt to change implicit measures in these domains will not consistently change behavior.
These results also produce a challenge for researchers who seek to understand the nature of human cognition because they raise new questions about the causal role of automatically retrieved associations. The results of the current meta-analysis do not lend themselves to a single interpretation. To better understand what the results mean, future research should innovate with more reliable and valid implicit, explicit, and behavioral tasks, intensive manipulations, longitudinal measurement of outcomes, heterogeneous samples, and diverse topics of study.

These innovations may yet reveal stronger evidence for the causal importance of automatically retrieved associations. It would not be the first time that the conclusions of a review were overturned by later advances. Following Wicker’s (1969) review showing a weak correlation between explicit attitudes and behavior, better measurement and theory revived the relevance of attitudes for understanding thought and action. As they did in response to Wicker, we hope that researchers take our findings as a challenge to improve theory and method and advance our understanding of human cognition.
META-ANALYSIS OF CHANGE IN IMPlicit MEASURES

References

* Contains data used in the meta-analysis.


http://dx.doi.org/10.1016/j.tics.2011.01.005


http://dx.doi.org/10.1002/bimj.201000151


* Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients’ approach bias for alcohol and improves treatment outcome. *Psychological Science, 22*, 490-497.


