

A META-ANALYSIS ON THE MALLEABILITY OF AUTOMATIC GENDER STEREOTYPES

Alison P. Lenton
University of Edinburgh

Martin Bruder
Cardiff University

Constantine Sedikides
University of Southampton

This meta-analytic review examined the efficacy of interventions aimed at reducing automatic gender stereotypes. Such interventions included attentional distraction, salience of within-category heterogeneity, and stereotype suppression. A small but significant main effect ($g = .32$) suggests that these interventions are successful but that their scope is limited. The intervention main effect was moderated by publication status, sample nationality, and intervention type. The meta-analytic findings suggest several issues worthy of further investigation, such as whether (a) other categories of intervention not yet identified or tested could be more effective, (b) suppression necessarily produces ironic effects in automatic stereotyping, (c) various indirect measures are differentially sensitive to stereotype change, and (d) automatic stereotypes about men differ in their malleability from those about women.

Gender is one of the most—if not the most—biologically primitive and important social categories (Kurzban, Tooby, & Cosmides, 2001). This would explain why it is the first social category that humans are able to discriminate (as early as 9 months of age; Leinbach & Fagot, 1993) and, consequently, why gender-related stereotypes are among the first stereotypes that humans develop (as early as age 2; Hill & Flom, 2007). Furthermore, stereotypes of men and women are complementary in a way that is unlike most other contrasting social categories (e.g., unlike Black vs. White ethnic groups; Glick & Fiske, 1996, 2001a). This between-group complementarity contributes to the maintenance of gender inequality, given that the distinct roles are perceived by many to be both natural and fair (Jost & Kay, 2005). Given their cultural embeddedness and seeming innateness, gender stereotypes can be particularly pernicious. To the extent that gender stereotypes impede men's and women's

progress or artificially limit their choices, it is important to understand if and how they might be counteracted. To that end, the present meta-analysis examines the efficacy of interventions aimed at reducing automatic gender stereotypes.

We focus on automatic stereotypes (i.e., those that are unintended—the respondent is either unaware of the assessed construct or unable to implement a particular response strategy; see Blair, 2002) because dual-system models of mental representation (Chaiken & Trope, 1999; Sloman, 1996; Smith & DeCoster, 1999) typically argue that automatic (vs. controlled) processes are relatively more resistant to change. Nevertheless, social psychological evidence for the malleability of automatic intergroup attitudes more generally has been accumulating in the past 10 or so years (see Blair, 2002, for a review). For example, with respect to gender, Blair, Ma, and Lenton (2001) reported that imagining a strong woman led to weaker automatic gender stereotypes than imagining a Caribbean vacation. Similarly, participants in another study (Steffens, Günster, & Hoffmann, 2005) were instructed to consider potential job applicants who were either counterstereotypical (i.e., an agentic female or a communal male) or stereotypical (i.e., a communal female or an agentic male). Participants in the former condition showed weaker automatic gender stereotypes as compared to those in the latter condition.

But what counts as change? Recently, Gregg, Seibt, and Banaji (2006) argued that researchers need to consider this continuum more carefully. For example, for interventions aimed at reducing automatic stereotypes to be considered truly effective, by how much should they reduce stereotypes? To reach this conceptual clarification, it

Alison P. Lenton, Department of Psychology, University of Edinburgh; Martin Bruder, Department of Psychology, Cardiff University; Constantine Sedikides, Department of Psychology, University of Southampton.

We thank Jamie DeCoster and Charles Bond for their assistance with our statistical queries, as well as the primary researchers who provided us their data for this meta-analysis. The research reported in this article was supported by Economic and Social Research Council grant #RES-000-22-0253.

Address correspondence and reprint requests to: Alison Lenton, 7 George Square, Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, Scotland, United Kingdom. E-mail: a.lenton@ed.ac.uk

would be helpful for researchers to know the degree of malleability of automatic stereotypes that has been empirically observed in intervention studies. Accordingly, we assessed meta-analytically the overall success of attempts to reduce automatic gender stereotypes. Indeed, providing an estimate of the mean success of attempts to reduce automatic gender stereotypes was the main goal of this meta-analysis; the search for moderators was another.

Before addressing these goals statistically, we first describe the model of stereotypes to which we adhere. In accordance with connectionist models (Smith & Conrey, 2007; Smith & DeCoster, 1998, 1999), we understand stereotypes as “‘states’ not ‘things’” (Smith & Conrey, 2007, p. 247). On the basis of this view, it might be construed as misleading for us to suggest that a stereotype could be “reduced” because this suggestion seems to imply that stereotypes are stable internal structures. Instead, connectionist models propose that stereotypes are quite elastic and, thus, any individual could hold an infinite number of representations of a social category’s members, when viewed across time and place. This is because a stereotype is a pattern of activation that, at a given point in time, is jointly determined by *current input* (i.e., the context) and the *connection weights* of the underlying network. These weights are incrementally updated over extended periods of time, as the individual encounters stimuli; updating of the connection weights is equivalent to learning. Thus, stereotypes are not static notions that people carry around in their heads no matter where they go; instead, the exact form that a stereotype takes depends both on people’s prior experience and on the judgment context in which they find themselves. For example, a person’s stereotype of women will likely differ if she is attending a conference alongside the top 100 businesswomen in the world, as compared to visiting a friend in the maternity ward of the local hospital. Consequently, when we suggest that there may be interventions that can successfully “reduce” automatic stereotypes, we mean to imply that these interventions, as (part of) current input, may produce an output pattern that is less consistent with traditional gender stereotypes than the pattern of activation that would emerge with more standard (stereotype-consistent or stereotype-irrelevant) input. In other words, asking people to imagine a “strong woman” prior to completing a measure of implicit gender stereotypes is likely to yield a less traditional stereotype than asking people to imagine a “weak woman” or a “Caribbean vacation” (Blair et al., 2001).

In light of the above, we make no strong theoretical claims about the longevity of the impact of any stereotype-reduction intervention, except to say that the intervention would likely lead to updating the connection weights. Because learning is a slow process, however, a single experience with a stereotype-reduction intervention is unlikely to change the connection weights to any substantial degree. Given that the vast majority of primary studies investigate stereotype change within single experimental sessions and

without repeated interventions, our meta-analysis should be viewed as examining malleability in current output activation patterns rather than in underlying connection weights.

Returning to the aims of this meta-analysis, in addition to providing an empirical effect size estimate of the relative power of stereotype-reduction interventions or, conversely, the relative inflexibility and resistance of automatic stereotypes to such interventions (Gregg et al., 2006, Studies 3–4), this meta-analysis may help to refine theorizing about automaticity and stereotyping more generally. The overall results will offer an indication of the general degree to which current input can—at least in the short term—override the default pattern of activation built up by the slow-learning system (Smith & Conrey, 2007; Smith & DeCoster, 1999). Again, connectionist models argue that output is a combination of both current input and the underlying connection weights, implying that the effects of a single instantiation of a stereotype-reduction intervention would be moderate at best. Our meta-analysis will provide a first quantification of the size of this effect.

Potential Moderators of the Effectiveness of Gender Stereotype-Reduction Interventions

We investigated seven potential moderators. The first three of these (i.e., intervention method, intervention specificity, type of indirect measure) describe the nature of the intervention or the automatic stereotyping measure used and, therefore, have theoretical implications for models of automatic stereotyping. The remaining four moderators (i.e., nationality of sample, gender composition of sample, publication status, sex of first author) refer to sample characteristics and publication features.

Intervention method. Researchers have examined the utility of a variety of interventions for changing automatic attitudes. These interventions range from manipulating experimenter race (Lowery, Hardin, & Sinclair, 2001) to instructing participants to see the world through the eyes of an elderly man (Galinsky & Moskowitz, 2000). In an attempt to organize this literature, Blair (2002) proposed five intervention categories: (a) Motivation (personal or social), (b) Stereotype reduction strategies, (c) Attentional focus, (d) Context cues, and (e) Characteristics of the target(s). However, as Table 1 shows, research on interventions that aim to reduce automatic gender stereotypes does not represent all five categories. Thus, we offer what we hope will be a productive alternative to intervention classification in the domain of automatic gender stereotypes.

In particular, we assigned each intervention to one of three categories (see Figure 1 for a summary of these intervention methods). The first, or our own category “A” interventions, distracts or redirects perceivers’ attention prior to category activation. The rationale behind this intervention

Table 1
 Characteristics of Studies Included in Meta-Analysis Testing the Malleability of Automatic Gender Stereotypes

Publication, study no.	Publication status	Sex of first author	Nationality of sample	Intervention specificity	Type of intervention ^a	Indirect measure	Percentage of male/female participants	Sample size ^b	Effect size (Hedges's <i>g</i>)	SE of <i>g</i>
Blair & Banaji (1996) Study 3	journal	female	US	both	heterogen	priming	37/63	70	1.53*** ^c	.27
Blair, Ma, & Lenton (2001)	journal	female								
Study 1			US	female	heterogen	IAT	40/60	39	.98**	.34
Study 2			US	female	heterogen	IAT	32/68	79	.67**	.23
Study 4			US	female	heterogen/suppress ^d	GNAT	32/68	102	.01	.22
Study 5			US	female	heterogen	DRM ^e	28/72	127	.52**	.18
Boccatto, Corneille, & Yzerbyt (2006)	unpublished	male								
Study 1			Belgium	both	suppress	priming	20/80	35	.21	.34
Study 2			Belgium	both	suppress	priming	20/80	44	.15	.30
Boccatto, Corneille, Yzerbyt, & Wittenbrink (2007)	unpublished	male								
Study 3			Belgium	female	suppress	priming	not available	48	-.05	.29
Carpenter (2001) Study 2	unpublished	female	US	both	heterogen	IAT	50/50	117	.43*	.19
Dasgupta & Asgari (2004) Study 1	journal	female	US	female	heterogen	IAT	0/100	72	.56*	.24
Study 2			US	female	heterogen	IAT	0/100	52	.90**	.29
Goodwin & Smoak (2007) Study 1	unpublished	female	US	both	heterogen	IAT	60/40	88	.21	.21
Häcker, Meyer, & Quinn (2007)	unpublished (conf. pres.)	female								
Study 1			UK	both	-f	cued recall	25/75	61	-.98*** ^g	.27
Liberman, & Förster (2000) Study 3	journal	female	US	female	suppress ^d	trait term production	47/53	45	-.20	.32
Macrae, Bodenhausen, Milne, Thorn, & Castelli (1997)	journal	male								
Study 1			UK	female	distract	LDT	50/50	32	.64†	.36
Study 2			UK	female	distract	LDT	0/100	32	.59	.36
Nodera & Karasawa (2005) Study 1	journal	female	Japan	female	distract	LDT	100/0	50	.36	.29

(continued)

Table 1 (continued)

Publication, study no.	Publication status	Sex of first author	Nationality of sample	Intervention specificity	Type of intervention ^a	Indirect measure	Percentage of male/female participants	Sample size ^b	Effect size (Hedges's <i>g</i>)	SE of <i>g</i>
Nosek & Banaji (2002)	unpublished (conf. pres.)	male								
Study 2			US	both	distract	GNAT	50/50	74	.27	.23
Steffens, Günster, & Hoffmann (2005)	unpublished	female								
Study 1			Germany	female	heterogen	IAT	33/67	143	.05	.17
Study 2			Germany	female	heterogen	IAT	23/77	192	.02	.14
Study 3			Germany	female	heterogen	IAT	33/67	144	.37*	.17

^aHeterogen = confrontation with heterogeneity within gender groups; suppress. = instruction to suppress stereotype expression; distract. = distraction or redirection of attention.

^bThe reported sample size might differ from the total sample size reported in the paper because (a) not all experimental groups were relevant to our analysis, (b) individual participants were not entered into the relevant analysis.

^cDue to its outlier status, this effect size was adjusted to $g = 1.13$ for all further analyses.

^dTwo dependent effect sizes were documented for this study. The average of these effects is reported here.

^eDRM = Deese-Roediger-McDermott false memory paradigm (Roediger & McDermott, 1995).

^fThis study used a cognitive load manipulation during the encoding phase of a memory task and, as such, it did not fit clearly into any of our categories. See footnote 2.

^gDue to its outlier status, this effect size was adjusted to $g = -.42$ for all further analyses.

* $p < .10$. ** $p < .05$. *** $p < .001$.

category is that a low level of engagement with the stimulus category would lead to little—if any—stereotype activation as compared to a higher level of engagement with the stimulus category. For example, in the context of a lexical decision task (LDT), participants in one study were shown digitized photos of women and household objects, some of which contained a white dot. The participants then either had to detect the dot's presence or decide whether the photograph contained an animate versus inanimate object (Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997). Those searching for the white dot supposedly had a lower level of engagement with the category stimulus than those judging whether the target was animate or not, and thus, they should be less likely to show gender stereotype activation in the subsequent implicit gender stereotyping task.

The second intervention type, or category "B" interventions, depends upon the existence of heterogeneity within the activated stereotype. Our research (Lenton, Sedikides, & Bruder, 2009) shows that representations of social categories can contain both stereotype-consistent and stereotype-inconsistent information at the same time. Interventions in this category may activate the representation, but emphasize a particular stereotype-inconsistent aspect of it. For example, before they completed a gender/leadership Implicit Association Test (IAT), participants in one study were given descriptions of either successful businesswomen or the origin and use of flowers (Dasgupta & Asgari, 2004). Thus, although a general representation might consist of relatively more stereotype-consistent depictions of women, the current input—"successful businesswomen"—brings the stereotype-inconsistent depictions to the fore.

The third type, or intervention category "C," is intended to prevent or inhibit stereotype expression, but not necessarily stereotype activation. For example, one experiment first trained participants to either say "yes" when they were presented with gender-stereotypical combinations of photos and words (e.g., a male photo paired with a male stereotype-consistent word) or to respond with "no" when they were presented with such combinations; following this procedure, participants completed a gender-priming task (Boccatto, Corneille, & Yzerbyt, 2006). As a result of this training, participants tried to suppress their general gender stereotypes when they encountered the subsequent priming task.

To summarize, category A interventions preclude or interfere with initial category and, thus, stereotype activation. Conversely, Category B and C interventions permit the category/stereotype to become activated and potentially guide further judgment. Category B and C interventions are distinct from one another, however, in terms of their focus of attention: Category B interventions direct perceivers' attention toward a particular aspect of the stereotype (i.e., the counterstereotypical aspect or subtype), whereas category C interventions activate the stereotype broadly, focusing perceivers' attention only on prevention or

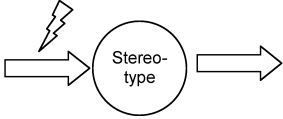
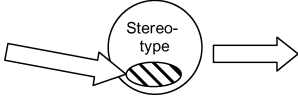
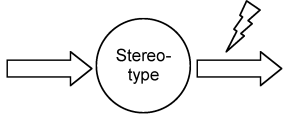
Intervention category	Process	Example
A	 Inhibit stereotype activation	Asking participants to focus on a white dot while they encounter stereotype-relevant material and before they complete an implicit measure of stereotypes.
B	 Emphasize stereotype heterogeneity by activating stereotype-inconsistent aspects of the category representation	Instructing participants to imagine a strong woman exemplar before completing a measure of implicit stereotypes.
C	 Prevent stereotype expression	Teaching participants to say “no” when encountering stereotypic stimulus combinations before measuring their implicit gender stereotypes.

Fig. 1. Characteristics and examples of intervention methods.

inhibition of its expression. Impression formation and person perception models (Brewer & Feinstein, 1999; Fiske, Lin, & Neuberg, 1999), in which category activation and attention constitute crucial and independent influences, support the distinctions we have made, as does research indicating that interventions that make the general category active (e.g., stereotype suppression) can produce ironic effects (i.e., the unintended consequence of increasing, rather than decreasing, subsequent stereotype activation; Macrae, Bodenhausen, Milne, & Jetten, 1994). Our meta-analysis, then, examines the relative effectiveness of these three intervention categories. We expect that, if any intervention category results in the temporary reversal of automatic gender stereotypes, it would be category B interventions because their current input is more likely than either category A or C interventions to activate counter-stereotypical subtypes.

Intervention specificity. Whereas some studies have sought to reduce automatic gender stereotypes in general (Blair & Banaji, 1996), others have focused exclusively on changing stereotypes about women (Dasgupta & Asgari, 2004). In this meta-analysis, we tested whether the specificity of the intervention (i.e., whether it focuses on stereotypes about women exclusively) is related to the effectiveness of the intervention. Research indicates that stereotypes of women are relatively more dynamic than stereotypes of men; stereotypes of women are perceived to have changed more during the last 50 years and are expected to change even more in the next 50 years (Diekmann & Eagly, 2000). Accordingly, we expected that interventions attempting to change beliefs about both men and women simultaneously would be less effective than those attempting to change beliefs about women only. As an example of

simultaneous belief-change interventions, participants in one study were instructed to expect a male name following a stereotypically feminine trait and a female name following a stereotypically masculine trait (Blair & Banaji, 1996). As an example of women-only belief-change interventions, in another study participants heard an aversive noise only after being presented with a negative female stereotypic word-pair, such as “female-weak” (Nodera & Karasawa, 2005). Note that no studies attempted to change stereotypes about men only. For more on this finding, see the Discussion below.

Type of indirect measure. Stereotyping measures are typically categorized as either explicit/direct or implicit/indirect, with little distinction made within each category. There is reason to believe, however, that indirect measures are not interchangeable. For example, debate surrounds the validity of Greenwald, McGhee, and Schwartz’s (1998) Implicit Association Test (Blanton, Jaccard, Gonzales, & Christie, 2006, 2007; Fiedler, Messner, & Bluemke, 2006; Nosek & Sriram, 2007). Indeed, the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001) was developed in response to one of the supposed shortfalls of the IAT, namely its inability to distinguish attitudes toward the group of interest versus attitudes toward a contrasting group. Additionally, research shows that apparently similar measures (e.g., LDT vs. conceptual priming) produce different results, with each having a unique relationship to explicit measures of the (supposedly) same construct (Wittenbrink, Judd, & Park, 2001). Still other research indicates that some indirect attitude measures are positively correlated (Cunningham, Preacher, & Banaji, 2001) and, thus, must assess the same construct to some degree. Therefore, in this meta-analysis, we examine whether

the effect of gender-stereotype-reduction interventions depends on the type of indirect measure employed.

Nationality of sample. Johnson and Eagly (2000) recommended that meta-analyses investigate, for generalizability purposes, the stability of effect-size estimates across geographic regions. Furthermore, research suggests that cultures vary in the extent to which they endorse gender stereotypes (Glick et al., 2000, 2004). It follows that stereotype-reduction interventions may be differentially effective across cultures.

Gender composition of sample. The majority of experimental psychology research relies on university convenience samples (e.g., introductory psychology students; Peterson, 2001; Sears, 1986). Female participants make up over half of these samples. Thus, research on automatic gender stereotypes may better reflect women's than men's gender-related representations. For example, Blair et al. (2001, Study 4) found that counterstereotypical mental imagery reduced automatic gender stereotyping only among female participants. These findings, together with research indicating that men are more likely than women to hold negative beliefs about women (Glick & Fiske, 1996, 2001b), bolster the utility of investigating whether the success of interventions to reduce automatic gender stereotypes depends on participant sex.

Publication status. A thorough and conservative meta-analysis includes both published and unpublished studies so as not to inflate the average effect size (Johnson & Eagly, 2000). Such inflation may result from what Rosenthal (1979) called the file drawer problem, where only significant findings tend to be published. We tested whether the *file-drawer problem* can account for effects of stereotype-reduction interventions.

Sex of first author. In a meta-analysis on sex differences in influenceability, Eagly and Carli (1981) reported that the size of the effect depended on author sex, such that male authors uncovered larger sex differences than did female authors. This finding has been interpreted as indicating that researchers tend to find or report results that are favorable to their own sex (Eagly & Wood, 1994; but see Hedges & Becker, 1986). To test for this possibility, we investigated the role of author sex in effect size magnitude.

Overview and Hypotheses

We conducted a meta-analysis of studies that focused on the reduction of automatic gender stereotypes. Our goal was to provide the first cumulative test of the potency of stereotype-reduction interventions or, conversely, the rigidity of automatic stereotypes. In view of connectionist models of mental representations, we expected that these interventions—as current input—would have a significant reductive effect on automatic stereotype output. However,

this effect would be moderate at best, given that existing connection weights also contribute to automatic stereotype output.

Furthermore, we sought to identify factors that moderate the effectiveness of such interventions. Based on previous theorizing and empirical results, we expected suppression-type interventions to be the least effective route to stereotype change. It was not clear, however, whether interventions involving attentional distraction or salience of heterogeneity would prove superior to the other. We also expected that interventions attempting to change beliefs about both men and women simultaneously would be less effective than those attempting to change beliefs about women only. Although we examined the impact of the type of indirect measure on automatic stereotype change, we did not have strong a priori hypotheses regarding which would be most or least sensitive, as researchers' understanding of the processing underlying them is limited.

Investigation of the role of sample nationality in the effects of stereotype-reduction interventions on automatic gender stereotypes was also exploratory, so our hypothesis here remained open. Given the predominance of female participants in most research on automatic gender stereotype change and the finding that, on average, men possess stronger and more negative stereotypes about women than women do, we expected that stereotype interventions would be more effective among women than among men. We anticipated that the effect size of unpublished studies would be lower than that of published studies, but that the file drawer problem would likely not fully account for the effect of stereotype-reduction interventions on automatic gender stereotypes. Finally, our investigation of the role of sex of the first author was exploratory: It was not clear what finding would be considered complimentary to the respective authors' gender group.

METHOD

Inclusion Criteria

For our meta-analysis, we selected studies that met the following criteria: (1) Stereotypes were investigated (i.e., conceptual associations) rather than prejudice or discrimination (Fiske, 1998); (2) Stereotypes concerned men and/or women in general, rather than male or female subgroups (e.g., elderly men); (3) An indirect measure of automatic gender stereotypes was used, where "indirect" was defined per Blair's (2002) conceptualization of automaticity; (4) The focus was on the malleability and, in particular, on the potential reduction of automatic gender stereotypes rather than on the general activation or even exacerbation of these stereotypes.

Literature Search

Database search. We searched the literature at the start of this project and again in November 2007 (near the close

of the project). As a first step in both searches, we submitted a combination of search terms to relevant online databases (PsycINFO, ISI Web of Knowledge, ERIC). A study needed to be located by all four search terms (corresponding to our four inclusion criteria) for it to be incorporated in the initial sample of studies for which titles and abstracts were screened:

- (1) (*stereotyp** OR *attitud** OR *prejud**) to locate stereotype-related research (allowing for imprecise categorizations by primary authors);
- (2) (*gender* OR *men* OR *women* OR *masculin** OR *feminin** OR *male* OR *female* OR *sex*) to limit the results to gender-related studies;
- (3) (*implicit* OR *automatic** OR *indirect* OR *unconscious** OR *nonconscious**) to locate studies investigating automatic processes; and
- (4) (*malleab** OR *chang** OR *influenc** OR *moderat** OR *reduc** OR *increas**) to locate studies focusing on change.¹

As an additional search criterion, we considered only studies published since 1989 because the assessment of automatic stereotypes became a major research endeavor in the 1990s, following the distinction between implicit and explicit racial attitudes (Devine, 1989). In our search occurring in November 2007, 549 PsycINFO entries met all four search criteria. This initial search, however, failed to identify a few relevant articles that we had gleaned informally from social psychological journals. Thus, we conducted a second search that relaxed the second criterion (gender), although, to keep results manageable, we used only the term *stereotyp** (and not *attitud** OR *prejud**) to satisfy our first criterion. This search resulted in 399 PsycINFO hits. We examined the titles and abstracts of all 798 publications (excluding duplicates) to identify studies that fulfilled our inclusion criteria.

Backward and forward search. After the database search, we conducted a backward search using the reference sections of all acceptable articles as well as the reference list of a narrative review on the malleability of automatic stereotypes and prejudice (Blair, 2002). Next, we carried out a forward search of PsycINFO and the Web of Knowledge to find studies that had since cited the identified papers or relevant references in the Blair (2002) article.

E-mail requests for support. The final step involved e-mailing (a) all first authors of relevant articles to inquire of additional studies they might have conducted and (b) authors of articles that met most, but not all, of our inclusion criteria to make a final determination regarding their relevance and to uncover unpublished work. We also requested relevant studies from the e-mail lists of the Society of Personality and Social Psychology, the European Association of

Experimental Social Psychology, and the social psychology section of the German Psychological Society.

Sample Characteristics and Recorded Variables

The final sample consisted of 13 research reports containing 21 independent effect sizes. For each effect size, we recorded the following features: (a) its publication status; (b) the nationality of the sample; (c) whether the male, the female, or both stereotypes were targeted by the intervention (intervention specificity); (d) the percentage of male and female participants; (e) the sample size; and (f) whether the intervention reversed the stereotype (for, although an effect size informs us if stereotyping is reduced or exacerbated, it does not by itself tell us whether an intervention effectively led to greater counterstereotyping than stereotyping). We also recorded the indirect dependent measure used to assess stereotype activation and change. The most commonly used measures were the IAT, the GNAT, sequential priming tasks (Fazio, Jackson, Dunton, & Williams, 1995), and LDTs (Macrae et al., 1994). Lastly, the first and second author independently coded the type of intervention used. In particular, we differentiated among three intervention categories (see Figure 1). The two raters initially agreed on 18 of the 21 categorizations. The categorizations for the three remaining effect sizes were resolved through discussion among the three authors of this article (a study corresponding to one of these three effect sizes was deemed uncategorizable with respect to our intervention classifications; see Table 1).

Effect Size Calculation

We used Hedges's *g* to assess effect size. In this measure, the mean difference between two groups is standardized by dividing it by the pooled standard deviation computed from both groups. Because our sample included a subset of all possible interventions designed to influence automatic attitudes (Blair, 2002), and we intended to ensure maximum generalizability of the findings, we used a random effects model in the overall integration of effect sizes and the examination of moderators (Hedges & Vevea, 1998). However, to represent more accurately the mean overall effect of our sample of studies, we also present the results of a fixed effects analysis. In all analyses, studies were weighted by the reciprocal of their variance (Hedges, 1994). We computed effect sizes and variance measures according to Johnson and Eagly (2000) and DeCoster (2004). We used Wilson's (2002) SPSS macros to compute the overall effect and to examine the impact of moderator variables.

RESULTS

Sample Descriptives

The sample of independent studies included in the meta-analysis was $k = 21$, with a total of $N = 1,646$. The mean

sample size was $n = 78.38$ and the median sample was $n = 70$ participants. Eighteen of the 21 studies showed an effect of the intervention in the expected direction, such that the group exposed to the stereotype-reduction intervention showed less automatic stereotyping than its respective control group. Eight of these effects were significant at $\alpha = .05$ (Table 1). Three studies revealed increased stereotyping in the intervention condition, with one of these effects reaching statistical significance. One study was based on a community sample (Dasgupta & Asgari, 2004, Study 1); the remainder were based on university students.

Outlier Detection

Prior to further analysis, we screened the data for possible outliers, using Huffcutt and Arthur's (1995) sample-adjusted meta-analytic deviancy (SAMD) statistic. The scree plot of the absolute value of the SAMD statistics revealed two outlier studies: the effect sizes observed by Blair and Banaji (1996, Study 3), $SAMD = 5.10$, and Häcker, Meyer, and Quinn (2007), $SAMD = 4.97$, were positive and negative outliers, respectively. One strategy for dealing with outliers is to exclude them from the meta-analysis. Alternatively, discrepant study effect sizes can be Windsorized and assigned a somewhat less extreme value (Lipsey & Wilson, 2001, p. 108). To be able to include these studies, we adjusted the two outlying effect sizes. To retain their relative extreme position, we assigned to them the value of the effect size of the next extreme study plus 0.5 standard deviations of the study sample ($SD/2 = .22$). For Blair and Banaji (1996), this meant adjusting the effect size from 1.53 to 1.20 for all further analyses. The effect size observed by Häcker et al. (2007) was adjusted from $-.98$ to $-.42$. These adjustments lowered the SAMD statistics of the outlying effect sizes to 3.70 and 2.84, bringing them within an acceptable range.²

Overall Effect of Interventions to Reduce Implicit Gender Stereotyping

The overall weighted mean effect was $g_{RE} = .32$ in the random effects analysis and $g_{FE} = .30$ in the fixed effects analysis, with a weighted standard deviation of .34. Both values were significant at $p < .0001$ (observed power $> .9999$) with 95% confidence intervals ranging from .18 to .46 for the random effects and from .21 to .38 for the fixed effects model. The observed range of effect sizes was $-.20 \leq g \leq .98$, not including the two outliers. Of the 20 studies for which it was possible to determine whether an intervention led to a reversal in stereotyping (i.e., the intervention evoked greater counterstereotyping than stereotyping), only four did so (Dasgupta & Asgari, 2004, Studies 1 and 2; Macrae et al., 1997, Studies 1 and 2). None of these reversals was statistically significant. As Table 1 indicates, two of the studies relied upon distraction interventions, and two relied upon exposure to within-category heterogeneity. Note that the study by Liberman and Förster (2000) could not be in-

cluded in the count because these authors did not measure counterstereotype activation.

Fail-safe numbers were calculated per Rosenberg (2005). In a fixed-effects model, the number of studies with null results (and a mean n equal to the present sample) that would be needed to reduce the overall effect to nonsignificance ($p > .05$) is 280. Rosenberg's (2005) estimates of fail safe numbers, which are less conservative than Rosenthal's (1979), suggest a number of 300 for the present analysis. Even a relatively large number of unpublished null findings would, therefore, not threaten the overall main effect, showing that interventions aimed at reducing automatic gender stereotypes have, on average, been successful. However, there was significant heterogeneity in the sample of effect sizes, $Q = 45.95$, $p = .0008$, suggesting the presence of moderators.

Moderator Analysis

Table 2 summarizes the results pertaining to moderators. Publication status, sample nationality, and type of intervention emerged as significant predictors of between-study heterogeneity, with no significant heterogeneity left within the respective groups. Published studies yielded a larger average effect size than unpublished studies, with the latter effect size being no different from zero. In addition, studies conducted with U.S. respondents yielded a larger average effect size than those conducted with European respondents; the latter effect was no different from zero. We found no support for a moderating effect of first-author sex or intervention specificity.

With respect to the type of intervention, those relying on attentional distraction or on increasing the salience of the heterogeneous nature of a gender stereotype (e.g., priming a counterstereotypical trait) had effect sizes significantly different from 0. Suppression interventions, on the other hand, did not differ from 0. Additionally, comparisons between the suppression and distraction interventions, $Q_B = 4.45$, $p = .035$, and between the suppression and heterogeneity interventions, $Q_B = 5.85$, $p = .016$, showed that distraction and heterogeneity interventions were both more effective than suppression at reducing automatic gender stereotypes; the effects of distraction and heterogeneity interventions were not significantly different from each other, $Q_B = .03$, $p = .855$. Thus, manipulations involving either distraction or directed attention to a particular (diverse) aspect of the stereotype had significant reductive effects overall and were reliably more powerful than those aimed at stereotype suppression. The latter, on average, had no effect.

The results for the type of indirect measure warrant additional attention. Although the nonsignificant omnibus test led us to abstain from conducting post hoc comparisons, the pattern of means and their associated significance levels nevertheless suggests that the GNAT, unlike the other indirect measures, may be impervious to or, perhaps, unable

Table 2
Analysis of Categorical Moderators Using a Random Effects Model

Moderator variable with respective levels	Q_B	Q_W	k	Hedges's g	SE of g	p of g
Publication status	8.76**	19.91				
Published		14.20	11	.55	.010	<.001
Unpublished		5.71	10	.14	.09	.124
First author	.16	20.95				
Female		19.02	15	.35	.09	<.001
Male		1.93	6	.28	.17	.101
Nationality of sample ^a	5.14*	20.14				
United States		13.89	11	.48	.10	<.001
Europe		6.25	9	.14	.11	.216
Intervention specificity	.10	20.74				
Both		9.15	7	.30	.14	.036
Female only		11.58	14	.36	.10	<.001
Type of intervention ^b	6.34*	16.06				
Distraction		.71	4	.43	.18	.020
Heterogeneity		14.55	11	.46	.09	<.001
Suppression		.81	5	.00	.16	.983
Indirect measure ^c	1.39	14.65				
IAT		7.55	9	.41	.11	<.001
GNAT		.30	2	.13	.24	.580
Priming		6.51	4	.40	.20	.042
LDT		.28	3	.51	.24	.035

Note. Q_B = between-groups Q statistic; Q_W = total within-groups Q statistic for moderator variable and separate Q statistic for each group. IAT = Implicit Association Test; GNAT = Go/No-Go Association Task; LDT = lexical decision task.

^aDue to insufficient sample size from non-U.S. and non-European countries, the study by Nodera and Karasawa (2005) had to be excluded from this analysis.

^bStudy 4 of Blair et al. (2001) reported effect sizes for both heterogeneity and suppression manipulations. Because these effect sizes used the same sample in the control condition and were thus partly dependent, only the effect size for the suppression condition was entered into this analysis.

^cWe only included indirect measures that were employed in at least two primary studies in this analysis.

* $p < .05$ (two-tailed). ** $p < .01$.

to detect change in automatic stereotypes. This null effect, however, is based on a very small sample and therefore potentially unstable.

We used a weighted least squares (WLS) regression, estimated via the method of moments, to compute the association between percentage of female participants and the effect size measure (see Steel & Kammeyer-Mueller, 2002, for an advocacy of WLS regression in this context). The regression provided no evidence for a relationship between the gender composition of the sample and the effect of stereotype-reduction interventions, $Q_{\text{Model}} = .19$, $p = .666$, $R^2 = .01$, $\beta = .10$. Thus, on the whole, these stereotype-reduction interventions were no more (or less) effective among women than among men.

Finally, we found that two significant moderators (publication status and sample nationality) were confounded, $\chi^2 = 5.05$, $p = .025$. Studies of U.S. samples were more likely to be published than studies of European samples. We entered these predictors simultaneously into a WLS regression to investigate whether they exert independent effects on effect size (Hedges, 1994). The combined moderators explained considerable heterogeneity in our sample, $Q_{\text{Model}} = 9.62$, $p = .008$, $R^2 = .33$, whereas the individual

beta weights were significant for publication status, $\beta = .45$, $p = .048$, and nonsignificant for sample nationality, $\beta = .21$, $p = .362$. Thus, publication status provides the larger contribution to variation in effect size.

DISCUSSION

The results of our meta-analysis show that interventions aimed at reducing automatic gender stereotypes have been successful overall, although the average effect size is small (Cohen, 1988). Automatic attitudes are indeed malleable and susceptible to some forms of single-session interventions (Blair, 2002). At the same time, however, the size of the effect indicates that interventions do not meet with unmitigated success. In particular, the interventions studied usually failed to reduce automatic stereotyping to zero and do not give rise to reliable counterstereotypic responding (Gregg et al., 2006). Whether there are substantial boundaries to the malleability of automatic responding and/or whether researchers have not yet identified the most powerful means for automatic stereotyping reduction remains unclear. Although our study sample did not contain interventions that manipulate participants'

motivations, it did include presumably potent interventions, such as distraction (minimal category activation) and exposure to counterstereotypical information. Thus, there is likely a limit on the degree to which automatic responding can be influenced by a single experience with a stereotype-reduction intervention.

Both publication status and sample nationality significantly moderated the effect of interventions on automatic gender stereotypes, such that published studies had a larger average effect size than unpublished studies, and studies using U.S. participants had a larger average effect size than those using European participants. There are several potential explanations for the latter finding. Perhaps gender stereotypes in these geographic regions differ in terms of their strength or content. Currently available implicit measures—especially those relying on semantic priming—may not be as valid outside the United States, as most have been developed with respect to North Americans' attitude and belief structures. It is also possible that particular interventions are more or less successful in one geographic region or another. Future research ought to investigate systematically the cross-cultural generalizability of implicit measures and stereotype interventions.

Publication status and sample nationality were correlated, however, and a subsequent multiple regression analysis revealed that publication status was the stronger predictor, with sample nationality falling to nonsignificance when controlling for publication status. Although these results indicate that small or nonsignificant effects are less likely to be published, they are not indicative of the worst-case file drawer problem, whereby the true effect size equals zero but the believed effect size is greater than zero. This is because we determined that at least 280 nonsignificant effects would be needed to revise our conclusion that automatic stereotype-reduction interventions are somewhat successful. At the same time, however, our results indicate that consideration only of published studies would lead to an overestimation of the success of stereotype-reduction interventions: The true success of these interventions is more modest than the published studies suggest.

The findings also indicate that some methods may be more (or less) effective than others. In particular, explicitly advising people to “just say no” (Boccatto et al., 2006) or to suppress their gender stereotypes (Blair et al., 2001, Study 4) does not result in a reduced automatic stereotype effect. These findings are important, as such campaigns are arguably among the most public and common types of interventions aimed at reducing unequal treatment of people. Contrary to other research (Macrae et al., 1994), however, we did not find that this particular intervention produced an *ironic effect*, whereby stereotypes were made more accessible following suppression (e.g., where someone might think even more about “women being homemakers” after trying to suppress this particular stereotypic image).

It is interesting to speculate on the observed lack of difference between the effectiveness of the distraction and

heterogeneity stereotype reduction interventions. One possibility is that the processes that mitigate automatic stereotyping in each intervention are unique, yet equally effective. From this perspective, we might advise equality campaigners either to (a) invent ways to distract individuals from processing information about a social category in an elaborate manner immediately prior to making a judgment about members of that category, or (b) instruct individuals to “think counterstereotypical thoughts” about category members before making judgments about them. Obviously, both recommendations are impractical to some extent, with the former likely to be especially difficult to implement outside the laboratory. In any case, before we can make any recommendations, it is necessary to point out that the automatic stereotyping measures were not randomly distributed across each type of intervention: Three out of the four distraction interventions were assessed with an LDT, and none of the heterogeneity interventions were assessed using this same measure. In fact, the method of measurement overlapped for just one study each (the GNAT; Blair et al., 2001; Nosek & Banaji, 2002). And when we compare the effect of heterogeneity (i.e., not averaged with suppression: Hedges' $g = .07$) to that of distraction on this measure (Hedges's $g = .27$), we find the effect of the latter to be nearly four times that of the former, suggesting—perhaps—that distraction-type interventions may ultimately be more effective at reducing automatic stereotypes than those that try to make counterstereotypes salient.

The findings also indicate that some methods of measuring stereotype change may be either less sensitive or, conversely, more automatic than others. In particular, the GNAT, unlike the other measures, did not show any overall effect of stereotype-reduction interventions. One potential explanation is that the GNAT was the only measure in the analysis to control for a possible shift in participants' response criterion, and this shift has been offered as an alternative explanation (vs. implicit associations) for the IAT effect (Brendl, Markman, & Messner, 2001). Blair et al.'s (2001) results contradict such an explanation, however, as one study (Study 5) used another measure that precludes the possibility of a response shift, and it showed significantly reduced automatic gender stereotypes. A second unique feature of the GNAT is that it does not require the use of a contrasting category of a similar level of abstraction (Nosek & Banaji, 2001). Further inspection of the methodology of the two GNAT studies reveals, however, that both relied on the male-contrasting category; thus, in practice, the GNAT was not so unique. Finally, research indicates that the internal consistency of the GNAT is low, both on average ($r = .20$, for the signal-detection version of the GNAT; Nosek & Banaji, 2001) and when compared to the internal consistency of other implicit measures (Nosek, Greenwald, & Banaji, 2007). Thus, it may simply be that the GNAT is insufficiently reliable to measure responsiveness to the interventions. Further research is needed with the GNAT to determine if and why this measure is different in

terms of its ability to capture or, alternatively, be resistant to stereotype malleability.

Neither the sex of author nor the sex composition of the sample contributed to variation in effect size. We can thus conclude that—at least in the domain of automatic gender-stereotype malleability—there is no evidence that authors find or report results complimentary to their own sex. In addition, men were no more (or less) susceptible to influence attempts than were women, even if these groups possessed (on average) a different starting point in terms of their beliefs about women (Glick & Fiske, 1996, 2001b). This finding suggests that belief strength does not moderate the effectiveness of stereotype-reduction interventions, although more direct evidence relevant to this interpretation is needed.

Our findings suggest that, whether the intervention aims to change only stereotypes about women or whether it aims to change gender stereotypes more generally, interventions may be equally effective. However, at this stage, it is still not possible to determine conclusively whether the male and female stereotypes are equally susceptible to interventions because few researchers have attempted to alter only the male stereotype. This finding in itself lends support to Miller, Taylor, and Buck's (1991) contention that men are perceived to be the normative category and women a deviation from this norm. We urge researchers to take up the challenge of seeking to determine whether male stereotypes are as susceptible to stereotype-reduction interventions as are female stereotypes or gender stereotypes more generally. Not only would this research serve to ameliorate a possible bias in our field, but it may help explain why the male role is perceived to have changed less over the last 50 years (Diekmann & Eagly, 2000), and it also may—albeit indirectly—provide support for our contention that the male stereotype is less heterogeneous than the female stereotype (Lenton, et al., 2009). Furthermore, given that men are, on average, liked less than are women (Eagly, Mladinic, & Otto, 1991; Rudman & Goodwin, 2004), it certainly seems there is ample scope for improving people's beliefs about and expectations of men.

Finally, our meta-analytic findings call attention to additional areas of research. There is a lack of studies investigating the duration of automatic gender stereotype change. Only one study in our sample (a quasi-experiment; Dasgupta & Asgari, 2004, Study 2) examined stereotype change beyond a single-session experiment. Again, connectionist models (Smith & Conrey, 2007; Smith & DeCoster, 1998, 1999) maintain that learning is a slow process and, as a result, a single experience with a stereotype-reduction intervention is unlikely to change the connection weights to any substantial degree, let alone for a lengthy period of time after the stereotype-reducing current input is removed. Future researchers would, therefore, be well advised to systematically investigate whether repeated exposure to a similar intervention reliably changes connection weights. One possibility is that, even if two intervention

methods are similarly successful in changing current output (e.g., distraction and heterogeneity interventions), they might be differentially potent in changing underlying connection weights over time. In particular, heterogeneity may be more effective over a longer time period. More research is also needed on how motives (be it self-motives or social motives; Blair, 2002; Sedikides & Strube, 1997) moderate automatic gender stereotypes. Finally, nearly all research on this topic has been conducted with young adults. It is conceivable that older individuals' stereotypes are more resistant to interventions such as those described in this article because single learning experiences should become less powerful over time relative to prior learning.

Conclusions

This meta-analysis demonstrates that interventions aimed at reducing automatic gender stereotypes have been successful on the whole, if not wholly successful, as these interventions were found to have a stable but small effect. The present findings also highlight several areas in need of additional research, including whether other categories of intervention could be more effective, if and when stereotype suppression results in ironic effects in automatic measures of stereotyping, if and how the GNAT is distinct from other indirect measures, and whether the male stereotype is as susceptible to reduction interventions as is the female stereotype. Our meta-analysis provides a clear picture of what research into the malleability of implicit gender stereotypes has revealed thus far and a solid footing on which to base future research.

Initial submission: February 18, 2008

Initial acceptance: July 31, 2008

Final acceptance: September 24, 2008

NOTES

1. Following the suggestion of an anonymous reviewer, we later included *context** in this search term to also identify studies that investigated contextual effects on automatic gender stereotypes. This, however, did not result in the identification of any additional relevant effect sizes.
2. The methods used by Blair and Banaji (1996) provide one clue as to this study's unusually large effect: In addition to receiving different interventions, participants in the control and experimental conditions also encountered different stimulus material in the dependent measure. In particular, participants in the experimental (vs. control) condition were presented with more counterstereotypic prime-target pairs. Arguably, this enhanced the ease with which participants could implement their strategy.

As indicated by our inability to assign Häcker et al.'s (2007) manipulation to an intervention type, the nature and potential effect of their manipulation were ambiguous. On the one hand, their manipulation of cognitive load was similar to a distraction manipulation and thus might have contributed to reduced automatic gender stereotyping (per Gilbert & Hixon, 1991). On the other hand, this distraction occurred during the encoding

phase of a memory task in which participants read both gender stereotype-consistent and -inconsistent sentences and, as such, the semantic processing of the material means that stereotypes could conceivably have become activated. The results indicate that the latter is likely to have been the case, but we based our inclusion of the study in this meta-analysis on theoretical, not empirical, grounds.

We also conducted all analyses without Winsorizing these two studies. The overall effects were virtually unchanged ($g_{RE} = .32$, $g_{FE} = .29$). The descriptive patterns for the moderator analyses were highly similar and significant moderator effects were identified for the same variables.

REFERENCES

- *References marked with an asterisk indicate studies included in the meta-analysis.
- Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, 242–261.
- *Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*, 1142–1163.
- *Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*, 828–841.
- Blanton, H., Jaccard, J., Gonzales, P., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*, 192–212.
- Blanton, H., Jaccard, J., Gonzales, P., & Christie, C. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology*, *43*, 399–409.
- *Boccatto, G., Corneille, O., & Yzerbyt, V. (2006). *Just say no: Effects of training in the negation of non-stereotypic associations on stereotype activation*. Unpublished manuscript, Université Catholique de Louvain, Belgium.
- *Boccatto, G., Corneille, O., Yzerbyt, V., & Wittenbrink, B. (2007). *Do not think of trait activation as stereotype activation! The reality of post-suppressional rebounds on stereotyping remains speculative*. Unpublished manuscript, Université Catholique de Louvain, Belgium.
- Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 760–773.
- Brewer, M. B., & Feinstein, A. (1999). Dual processes in the representation of persons and social categories. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 255–270). New York: Guilford.
- *Carpenter, S. J. (2001). Implicit gender attitudes. (Doctoral dissertation, Yale University). *Dissertation Abstracts International*, *61*(10-B), 5619.
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science*, *12*, 163–170.
- *Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*, 642–658.
- DeCoster, J. (2004, September 19). *Meta-analysis notes*. Retrieved April 1, 2007, from <http://www.stat-help.com/notes.html>
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18.
- Diekmann, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*, 1171–1181.
- Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-type communications as determinants of sex difference in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin*, *90*, 1–20.
- Eagly, A. H., Mladinic, A., & Otto, S. (1991). Are women evaluated more favorably than men? *Psychology of Women Quarterly*, *15*, 203–216.
- Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 485–500). New York: Russell Sage Foundation.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.
- Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the “I,” the “A,” and the “T”: A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*, 74–147.
- Fiske, S. T. (1998). Prejudice, stereotyping, and discrimination. In G. Lindzey (Ed.), *The handbook of social psychology* (pp. 357–411). New York: McGraw-Hill.
- Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 231–254). New York: Guilford.
- Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*, 708–724.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, *60*, 509–517.
- Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*, 491–512.
- Glick, P., & Fiske, S. T. (2001a). Ambivalent sexism. *Advances in Experimental Social Psychology*, *33*, 115–188.
- Glick, P., & Fiske, S. T. (2001b). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications of gender inequality. *American Psychologist*, *56*, 109–118.

- Glick, P., Fiske, S. T., Mladinic, A., Saiz, J., Abrams, D., Masser, B., et al. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*, 763–775.
- Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., et al. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, *86*, 713–728.
- *Goodwin, S. A., & Smoak, N. D. (2007). Implicit gender stereotype change: A social role perspective. Unpublished raw data.
- Greenwald, A. G., McGhee, D. E., Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.
- Gregg, A. P., Seibt, B., & Banaji, M. A. (2006). Easier done than undone: Asymmetry in the malleability of automatic preferences. *Journal of Personality and Social Psychology*, *90*, 1–20.
- *Häcker, C., Meyer, A., & Quinn, K. (2007, September). *Effects of cognitive load on online processing and memory of stereotype relevant information*. Poster session presented at the BPS Social Section conference, Canterbury, UK.
- Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.
- Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 14–50). Baltimore: Johns Hopkins University Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.
- Hill, S. E., & Flom, R. (2007). 18- & 24-month-olds' discrimination of gender-consistent and inconsistent activities. *Infant Behavior and Development*, *30*, 168–173.
- Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology*, *80*, 327–334.
- Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 496–528). New York: Cambridge University Press.
- Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: Consequences for specific and diffuse forms of system justification. *Journal of Personality and Social Psychology*, *88*, 498–509.
- Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*, 15387–15392.
- Leinbach, M. D., & Fagot, B. I. (1993). Categorical habituation to male and female faces: Gender schematic processing in infancy. *Infant Behavior & Development*, *16*, 317–332.
- Lenton, A. P., Sedikides, C., & Bruder, M. (2009). A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles*, *60*, 269–278.
- *Liberman, N., & Förster, J. (2000). Expression after suppression: A motivational explanation of postsuppressional rebound. *Journal of Personality and Social Psychology*, *79*, 190–203.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage.
- Lowery, B., Hardin, C., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*, 842–855.
- Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, *67*, 808–817.
- *Macrae, C., Bodenhausen, G. V., Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the activation of social stereotypes: The moderating role of processing objectives. *Journal of Experimental Social Psychology*, *33*, 471–489.
- Miller, D. T., Taylor, B., & Buck, M. L. (1991). Gender gaps: Who needs to be explained? *Journal of Personality and Social Psychology*, *61*, 5–12.
- *Nodera, A., & Karasawa, K. (2005). The inhibitive effect of punishment on stereotype activation. *Japanese Journal of Social Psychology*, *20*, 181–190.
- Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, *19*, 625–666.
- *Nosek, B. A., & Banaji, M. R. (2002, February). *The power of the immediate situation: Gender differences in implicit math attitudes*. Paper presented at the Conference for the Society of Personality and Social Psychology, Savannah, GA.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.
- Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales and Christie (2006). *Journal of Experimental Social Psychology*, *43*, 393–398.
- Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 250–261.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814.
- Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*, 464–468.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.
- Rudman, L. A., & Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, *87*, 494–509.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530.
- Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 209–269). New York: Academic Press.

- Sloman, S. A. (1996). The empirical case of two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.
- Smith, E. R., & Conroy, F. R. (2007). Mental representations are states, not things: Implications for explicit and implicit measurement. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 247–264). New York: Guilford.
- Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*, 21–35.
- Smith, E., & DeCoster, J. (1999). Associative and rule-based processing: A connectionist interpretation of dual-process models. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 323–338). New York: Guilford.
- Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*, 96–111.
- *Steffens, M. C., Günster, A. C., & Hoffmann, C. (2005). *Sugar and spice make everybody seem nice: Dissociating the influences of sex and gender in the ascription of leadership qualities*. Unpublished manuscript, University of Jena.
- Wilson, D. B. (2002, January 15). *SPSS macros for performing meta-analytic analyses*. Retrieved April 1, 2007, from <http://mason.gmu.edu/~dwilsonb/ma.html>
- Wittenbrink, B., Judd, C. M., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology*, *37*, 244–252.