Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: https://www.tandfonline.com/loi/hbas20

The Influence of Instructions to Correct for Bias on Social Judgments

Saera R. Khan , Tzipporah Dang & Andrea Mack

To cite this article: Saera R. Khan , Tzipporah Dang & Andrea Mack (2014) The Influence of Instructions to Correct for Bias on Social Judgments, Basic and Applied Social Psychology, 36:6, 553-562, DOI: 10.1080/01973533.2014.971157

To link to this article: https://doi.org/10.1080/01973533.2014.971157

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 36:553–562, 2014

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2014.971157

Routledge
Taylor & Francis Group

The Influence of Instructions to Correct for Bias on Social Judgments

Saera R. Khan, Tzipporah Dang, and Andrea Mack
University of San Francisco

We examined how instructions to correct for bias influenced judgments of a male target person whose behavior towards a female was either negative or ambiguous. Half of the female participants with egalitarian or traditional views about gender were instructed to correct for bias prior to reading the vignette. All participants rated his negative behavior unfavorably. In the non-instructed condition, participants with a traditional bias rated the ambiguous male behavior more favorably than participants with an egalitarian bias. However, in the instructed condition, this pattern was reversed. Results demonstrate that the evaluative implications of behavior can impact correction effects.

Social judgments reflect the combined influence of multiple sources of information, some of which are directly observable (e.g., skin color) and others that we must infer (e.g., intentionality). Previous work has shown that subjective factors, such as attitudes, personal beliefs, current mood, and social biases can influence social judgments (e.g., DeSteno, Petty, Rucker, Wegener, & Braverman, 2004; Lambert, Khan, Lickel, & Fricke, 1997; Petty, Priester, & Wegener, 1994; Wyer & Budesheim, 1987). Our reliance on subjective factors to form social judgments is especially likely under conditions of uncertainty, when directly observable information is scarce (e.g., Borgida & Howard-Pitney, 1983). For example, judging whether someone's ambiguous behavior, such as skydiving, is adventurous or reckless is influenced by the recent activation of these concepts (Higgins, Rholes, & Jones, 1977).

Understanding the factors that influence the formation and accuracy of social judgments is crucial because they shape our first impressions and how we respond to others. The consequences of misjudging can harm social relations and contribute to prejudice and discrimination. Fortunately, perceivers can adjust their judgments if they feel that they have been influenced by incorrect or inappropriate information. A correction effect is seen when a social judgment is revised in the opposite direction of a presumed bias, such that the person or situation would be judged either more favorably or unfavorably than if the correction did not occur (e.g., Sczesny & Kühnen, 2004; Strack & Mussweiler, 2001; Wegener & Petty, 1995). Biases are often not noticed unless perceivers already have metacognitive knowledge of their potential influence (see Petty & Wegener, 1993; Sczesny & Kühnen, 2004). However, previous work suggests that alerting people to their possible biases can influence the direction and magnitude of their judgments (Monteith, Deneen, & Tooman, 1996; Wegener & Petty, 1995, 1997; Wilson, Centerbar, & Brekke, 2002).

The goal of the current study is to better understand the circumstances under which the judgment process is more or less likely to reflect sources of gender bias and the extent to which we can correct for possible biases by drawing awareness to them. We examined how instructions to correct for bias influenced judgments of a male target person whose behavior toward a female was either clearly negative or ambiguous. We further explored how instructions to correct for bias impacted social judgments according to the strength and direction of women's gender beliefs (egalitarian vs. traditional).

MENTAL CORRECTION EFFECTS

Correction theories such as Wilson and Brekke's (1994) theory of mental contamination, the flexible-correction model (Wegener & Petty, 1997), and Fazio's motivation

Correspondence should be sent to Saera R. Khan, Department of Psychology, University of San Francisco, 2130 Fulton Street, San Francisco, CA 94117. E-mail: srkhan@usfca.edu

and opportunity as determinants (Fazio & Towles-Schwen, 1999) center on people's subjective theories about their own bias and its influence on judgment (see also Martin, 1986).[1] Perceivers decide if there is an undesirable influence contaminating their judgment if clear or objective criteria for measuring accuracy are not readily available. People then consult their own naïve theories of judgment to determine the extent and direction of bias that might have an impact on their judgments and adjust accordingly (Strack, 1992; Wegener & Petty, 1995).

Determining correction effects involves comparing magnitude or changes in the direction of judgment ratings between a control and experimental group (i.e., between those who were not given special instructions to correct for biases and those who were) and the direction of participants' respective bias. Undercorrection is observed when the magnitude in judgment decreases but the direction of the judgment remains consistent with the original bias (e.g., correction adjusts a negative rating of a target's behavior so that it is less extreme). Overcorrection occurs when the direction of the judgment is in the opposite direction predicted by the bias (e.g., correction adjusts a negative rating of a target's behavior so that it is evaluated favorably). A successful correction would occur if the particular bias did not predict the evaluation of the target person, meaning that regardless of the valence of the bias the target person was rated similarly. Although correction attempts are not always perfectly calibrated, people do seem to have an awareness of the direction but not always the magnitude of their bias.

One issue in need of closer scrutiny is the impact of instructing people to correct for bias prior to judgment (e.g., Petty & Wegener, 1993; Stapel, Martin, & Schwarz, 1998; Wegener & Petty, 1997). Some researchers advocate this practice as an effective tool in reducing discriminatory judgments (Fleming, Wegener, & Petty, 1999; Schuller, Kazoleas, & Kawakami, 2009; Sommers, 2006; Wilson & Brekke, 1994). However, research has shown that the type of correction instructions provided to participants produces differences in correction effects (Stapel et al., 1998).[2] In one such study, participants were asked to rate U.S. midwestern cities based on the desirability of their weather. Prior to these ratings, participants rated vacation spots in terms of weather desirability (i.e., a salient source of bias) or job satisfaction (i.e., a subtle source of bias). When participants were blatantly instructed that their prior ratings of vacation spots would influence their ratings of midwestern cities, participants adjusted their latter ratings regardless of whether the source of bias was salient or subtle. In this case, participants do not necessarily detect and correct for bias on their own but engage in correction because they have been explicitly instructed to do so. However, when participants were warned that bias (without explicitly identifying the source of bias) might influence their judgments of desirability of midwestern cities, participants corrected for their ratings only when they had rated vacation spots on the same dimension but not when the source of bias was subtle and unapparent to them. According to these researchers, subtle correction instructions that do not name the bias can cue participants to search for possible bias prior to making a judgment. Different instruction sets can modify the extent to which correction effects are influenced by contextual information.

In addition to instruction sets, judgments that involve rating target members on stereotypically relevant dimensions can also influence the correction process. For example, Sczesny and Kühnen (2004) demonstrated that even without correction instructions, awareness of the cultural stereotype that men are more competent leaders than women led perceivers to correct for these beliefs prior to judgment. A cognitive load manipulation was used in their study to demonstrate differences in rating men and women on leadership competence. Participants in the cognitive load condition rated men as more competent leaders than women, as expected from the cultural stereotype. However, participants in the nonmanipulated condition showed no differences in leadership ratings based on the gender of the target persons, resulting in a correction effect. Together, these studies demonstrate that correction effects are influenced by contextual stimuli, whether it be types of instruction sets or judgments on stereotypically relevant dimensions; cues within our social environment influence our awareness and correction of our biases in judgment (see Kunda & Spencer, 2003). Given that subtle correction instructions had a differential effect on the correction process depending on whether or not the source of bias was clear to participants, it is important to understand whether blatant or ambiguous social information about a target person can also influence the correction process by prompting participants to become aware of their own biases.

Most studies that have examined correction effects involve judgment of a target person with only categorical information provided (e.g., gender or race) and

[1]Several notable models exist accounting for correction effects in social judgments (e.g., Fazio and Towles-Schwen's, 1999, motivation and opportunity as determinants model; Petty and Wegener's, 1993, 1997, flexible correction model; Martin's, 1986, set/reset model; and Schwarz and Bless's, 1992, inclusion/exclusion model). Although the mental contamination model and flexible correction models are more heavily referenced here, similar predictions can be made with these aforementioned models.

[2]This study conducted by Stapel and colleagues has not been retracted: http:/c.ymcdn.com/sites/www.spsp.org/resource/resmgr/docs/nonretractionreportpspbinpre.pdf

very little else regarding the person's behavior, leaving a critical question unaddressed: Does the extent to which biases are susceptible to correction depend upon the ambiguity of the behavior being judged? We know that differences in social judgments rest on the availability of information about the target person and the perceiver's motivation to make an appropriate judgment (see Fiske, Lin, & Neuberg, 1999; Fiske & Neuberg, 1990). For example, the application of ingroup favoritism, a type of bias, is moderated by the evaluative implications of a target person's behavior (e.g., Marques, Robalo, & Rocha, 1992; Marques & Yzerbyt, 1988). When behavior is ambiguous, people are more likely to judge an ingroup member favorably but less likely to apply this bias when the ingroup member's behavior is blatantly negative (Khan & Lambert, 1998). If the ambiguity of the target's behavior plays a role in the extent to which our biases are relied upon for judgment, then this factor may also play an important role for when we correct for our biases.

Role of Prior Beliefs

Although theories on correction do not specify that nonegalitarian or prejudiced views are somehow especially vulnerable to correction, in an increasingly egalitarian society where public expressions of negative prejudice are frowned upon, it is reasonable to assume that correction instructions would have a differential effect depending on whether someone possessed an egalitarian or traditional bias. With respect to gender and race, research reveals the intriguing possibility that correction effects can also operate on egalitarian beliefs (Lepore & Brown, 2002; Olson & Fazio, 2004). For example, women are reluctant to label someone sexist especially if they perceive the behavior as unintentional or harmless (Swim, Scott, Sechrist, Campbell, & Stangor, 2003). This reluctance may stem from an intense desire to be "fair," which can undermine egalitarian concerns. In another set of studies, low- and high-prejudiced participants were primed with Black or White faces before judging an African American target person depicted in a photo. Participants corrected for both their favorable and unfavorable bias toward an African American target person if they were especially motivated to avoid any type of dispute about their racial attitude (Olson & Fazio, 2004). Of interest, low-prejudiced participants primed with Black faces judged the African American target more negatively than expected from their relatively egalitarian beliefs. Taken together, these studies suggest that, in addition to modulating the impact of prejudiced views on social judgments, egalitarian views are also susceptible to correction. The design of the present study will allow for a better understanding of whether bidirectional correction effects can be obtained only in priming paradigms with categorical information or if they can also be obtained when participants are presented with social information that requires their discernment for whether their bias might have an influence on their judgments.

Overview of Study

We tested whether instructions to correct for bias operated differently for participants with traditional or egalitarian gender bias. Participants were not explicitly instructed to correct for gender bias but were simply told to avoid bias when forming a judgment. The rationale for not providing explicit instructions for "which" bias to correct for is that it allows a stronger test for observing whether participants cue into the direction and magnitude of their bias prior to correction. The only gender-laden information provided was the first names of the conversants (i.e., Jim and Ann). We also examined how instructions to correct for bias influenced judgments of a male target person whose behavior toward a female was either clearly negative or ambiguous. In both the instructed and noninstructed conditions, a male target person makes comments to a female student about her academic performance that are either ambiguous (i.e., can be perceived as helpful or condescending) or blatantly negative.

It is important to clarify how we conceptualize "bias" in the present study. Definitions for "bias" differ across disciplines and even subdisciplines within psychology (for a review, see Hahn & Harris, 2014). In this study, bias refers to a tendency or preference for a particular worldview. In agreement with previous work, we assume that a correction effect occurs when a social judgment is revised in the opposite direction of a presumed bias, such that the person or situation would be judged either more favorably or unfavorably than if the correction did not occur (e.g., Sczesny & Kühnen, 2004; Strack & Mussweiler, 2001; Wegener & Petty, 1995). Although an argument can be made that egalitarian views represent a "biased" free perspective, our working assumption is that the particular "bias" operating is a preference for nontraditional roles and power structures that sensitizes one to situations suggestive of traditional displays of bias and discrimination. In contrast, a traditional bias is characterized by a preference for conventional social roles and the existing hierarchical power structures.

In the negative condition, we predicted that participants would rely more strongly on the blatancy of the comments as opposed to their bias to judge the conversation and male target person. Therefore, we predicted that all participants would rate the male unfavorably and that gender bias and the correction instructions would have little effect on their judgments (Hypothesis 1). In the ambiguous condition, however, both the direction and

magnitude of judgment were expected to differ by gender bias. The most dramatic correction effects were expected to occur in the ambiguous as opposed to the negative condition because gender bias would exert greater influence in rendering judgment in the former condition than in the latter condition. With no instructions to correct for bias, we predicted that women with an egalitarian bias should have a more unfavorable reaction to the conversation and the male target person relative to women with a traditional gender bias who would perceive ambiguous comments favorably. Therefore, in the ambiguous condition, women with a traditional gender bias will rate the male target more favorably than women with an egalitarian bias (Hypothesis 2). When instructed to correct, we predicted that women with a traditional gender bias would adjust their favorability ratings downward and those with an egalitarian bias would adjust their favorability ratings upward. As a result of correction, the difference in favorability ratings observed in the noninstructed condition should either be diminished or reversed (Hypothesis 3).

METHOD

Participants and Design

A total of 130 White American female undergraduate students at a public university participated in the experiment in exchange for partial course credit for their introductory psychology class. Participants' ages ranged from 18 to 22 years, with a mean age of 19 years ($SD = .63$). All were native English speakers. The experimental design was a 2 (comment type: ambiguous vs. negative) $\times$ 2 (instruction set: general vs. correction) fully crossed between-participant design with gender bias as a continuous variable.

Materials and Procedure

At the start of the study, participants were ushered into separate cubicles. Participants were randomly assigned to read an excerpt from a conversation between two students in which the male makes comments (ambiguous or negative) toward the female student. Half of the participants were instructed to read the transcript and form an impression (i.e., general instruction set), whereas the other half read correction instructions. Participants in the correction instruction condition were also provided with the following set of instructions:

> Important: When we are judging other people's behavior, previous research has shown that in order to arrive at an accurate appraisal, it is necessary that you be as logical and analytical as you can. This research has shown that part of being accurate means that you be particularly aware of any factor that might "bias" your answer and adjust for these biases in the most careful manner possible. This process requires much attention and effort so please work on each question as carefully as you can until you have arrived at the most accurate appraisal possible.

All participants were provided with a transcript of a conversation that ostensibly took place in our lab as part of a study examining "get acquainted" conversations. The dialogue provided very little information about the female conversant because our primary interest was in reactions toward the male target. In the dialogue, Jim asks Ann about her academic progress in her major classes. Ann reports making Bs and Cs in her classes. In the ambiguous condition, Jim remarks, "C's and B's . . . hmm . . . Have you thought about getting a tutor to help you out?" To ensure that the comments were indeed ambiguous, we adopted transcripts used in a published study examining ingroup favoritism effects (Khan & Lambert, 1998). In this prior study, male and female participants' reactions to the ambiguous comment depended on whether the target person was part of their ingroup. That is, participants rated the ambiguous comment more favorably when it was made by a fellow ingroup target member than when it was made by an outgroup member (Khan & Lambert, 1998). In the negative condition, Jim suggests that she change to an easier major. This suggestion was blatantly rude and insulting to Ann. The transcripts were identical prior to Jim's advice to Ann about her mediocre grades and Ann's vague response to his advice (i.e., "Well, I never really thought about it.") was identical across conditions.

After reading the transcript, participants provided their reactions toward the conversants and the conversation as a whole. The evaluation of the conversation ("What was your overall reaction to the conversation?") and the two conversants (e.g., "What was your overall reaction towards Jim?" and "How much empathy do you have for Jim?") was provided along a scale from -5 (*not at all favorable*) to $+5$ (*very favorable*). Following this, participants rated Jim and Ann along a scale ranging from 0 (*not at all*) to 10 (*extremely*) for each of the following traits*: rude, intelligent, friendly, honest, independent, polite, kind,* and *sexist*. These traits and questions related to emotional reactions to the conversation and conversants were chosen to capture the two universal dimensions by which we judge others: warmth and competence. A wealth of research has established that people judge others on the basis of perceived warmth and competence (for a review, see Fiske, Cuddy, & Glick, 2007). Last, they completed the Ambivalent Sexism Inventory (ASI; Glick & Fiske, 1996).

The ASI comprises two interrelated subscales (Hostile and Benevolent Sexism) that jointly represent

an ideological belief system that perpetuates traditional notions of hierarchical power and roles for males and females within a social system (see Glick & Fiske, 1996, 2001, 2011). For both genders, ASI is strongly related to individual differences in the need for cognitive closure (Kruglanski, 1990), and this relationship is mediated by the classic predictors of antiegalitarianism, social dominance orientation, and right-wing authoritarianism (Roets, Van Hiel, & Dhont, 2012; Sibley et al., 2009). Hostile sexism reflects a negative attitude toward women who challenge traditional notions of men's domination and power. Benevolent sexism is subtle and encompasses the view that women ought to be protected and cherished because of their intrinsic fragility. Although use of the term "ambivalent" for this construct has been mistaken by some researchers to mean two opposing subscales, Glick and Fiske (2011, 2012) clarified that the scales together reflect a single ideological belief system with respect to beliefs about gender. Recent research exploring the antecedents of ambivalent sexism reveals that women's beliefs about gender are informed by their overall worldview (Roets et al., 2012). Using the ASI, researchers found that the two ASI subscales tap into two distinct measures of antiegalitarian attitudes; the Benevolent subscale is predicted by the value for traditionalism and obedience as measured by the Right Wing Authoritarianism scale (Altemeyer, 1981), and the Hostile subscale is predicted by the endorsement of existing social hierarchies as measured by the Social Dominance Orientation scale (Pratto, Sidanius, Stallworth, & Malle, 1994). Combining the two subscales as a measure of gender bias ensures that both of these underlying dimensions of egalitarian bias are captured (see also Sibley, Wilson, & Duckitt, 2007). Cross-cultural studies demonstrate that these two subscales are distinct but highly correlated and do indeed aid in the justification and maintenance of gender inequality (Glick, 2006; Glick et al., 2000; Glick et al., 2004). Women scoring low on both forms of the ASI scale can be seen as rejecting traditional sex roles and are likely to view benevolent sexism as paternalistic rather than helpful. Table 1 contains correlations and reliability information for the subscales and overall scale. Analyses involving the effects of each subscale are footnoted and results from using the overall ASI scale are reported next.[3]

[3]Hierarchical regression analyses were performed using hostile and benevolent subscales as predictors. The main effects of advice type, instruction set, and hostile sexism scores were entered on the first step, all two-way interactions were entered on the second step, and the three-way interaction was entered on the third step. Results revealed a nonsignificant three-way interaction ($\beta = .33$, $p = .073$, $R^2 = .33$), $F(7, 114) = 7.95$, $p < .01$. Similar analyses using the Benevolent subscale as the predictor instead revealed the same pattern of results for the three-way interaction ($\beta = .27$, $p = .104$, $R^2 = .34$), $F(7, 114) = 8.23$, $p < .01$.

TABLE 1
Descriptive Statistics and Correlations Between Benevolent Sexism, Hostile Sexism, and Ambivalent Sexism Inventory and Ratings of Male Conversant

Scale	*1*	*2*	*3*	*4*
1. Benevolent Sexism subscale				
2. Hostile Sexism subscale	.42**			
3. Ambivalent Sexism Inventory	.84**	.85**		
4. Composite rating of reaction to male conversant	.08	.14	.13	
M	1.94	1.88	1.91	0
SD	.87	.88	.74	1.0
α	.80	.85	.87	.88

Note. $N = 125$ women for all analyses. Scores for Ambivalent Sexism Inventory, Hostile Sexism, Benevolent Sexism ranged from 0 to 5.
**$p < .01$.

RESULTS

Scoring

A principal components analysis with a varimax rotation was conducted to create a composite rating for overall reactions to the male conversant. The first factor captured 47% of the variance (eigenvalue = 5.16); then there was a large decline and bend in eigenvalues for the second factor (1.30) and third factor (1.11). A weighted factor score using a regression approach was created for the first factor and was based on participants' ratings on all items included in the analysis ($\alpha = .88$; for information on weighted factor scores, see Robins, Fraley, & Krueger, 2007). This approach allowed the creation of a composite explaining as much predictor variation as possible. This new composite was standardized and has a mean of 0 and a standard deviation of 1 (item values on the rotated component matrix: emotional reaction =.73, emotional reaction to Jim = .80, empathy for Jim = .79, rude = .28, intelligent = .60, friendly = .38, honest = − .10, independent =.44, polite = .58, kind = .48, and sexist = − .05). From this point on, references to the judgment of the male conversant are based on this composite.

Preliminary Analyses

Prior to testing the hypotheses described earlier in the article, manipulation effects and order effects were tested. All dependent variables were recoded along a 0-to-10 scale, and all negatively valenced items (e.g., rude and sexist) were reverse scored. A one-way analysis of variance was conducted to assess if the manipulation of comment type had its intended effect on participants' reaction to Jim and the conversation. Participants rated the conversation more unfavorably in the negative condition ($M = 3.55$, $SD = 1.77$) compared to the ambiguous

condition ($M = 5.17$, $SD = 2.03$), $F(1, 127) = 23.53$, $p < .001$. As expected, participants also reacted more unfavorably toward Jim in the negative compared to the ambiguous condition ($M = 2.30$, $SD = 2.05$ vs. $M = 5.12$, $SD = 2.99$), $F(1, 127) = 40.01$, $p < .001$. To evaluate the possibility that participants' reported ASI scores may have been influenced by their prior assignment to experimental conditions, a 2×2 analysis of variance (comment type and instruction set) with ASI scores as the dependent variable was conducted. No significant effect was observed as a function of instruction set. ASI scores for participants given correction instructions ($M = 1.89$, $SD = .70$) did not differ significantly from scores for participants given general instructions ($M = 1.93$, $SD = .78$), $F(1, 125) = .01$, $p = .965$. A significant main effect for comment was observed, $F(1, 125) = 4.29$, $p = .040$, $\eta^2 = .03$. ASI scores in the ambiguous comment condition were higher ($M = 2.06$, $SD = .71$) relative to scores in the negative comment condition ($M = 1.79$, $SD = .74$). Although there was a statistically significant difference by comment condition, the mean difference was very small (−.27) and, of importance, the interaction effect was not significant, $F(1, 121) = .06$, $p = .803$. In the ambiguous condition, the mean ASI score for participants in the general instruction set was similar ($M = 2.05$, $SD = .79$) to scores for participants in the correction instruction set ($M = 2.08$, $SD = .63$). The scores in the negative condition were also no different (general: $M = 1.81$, $SD = .78$; correction: $M = 1.77$, $SD = .72$). Thus, ratings in all four experimental conditions revealed that ASI scores were not influenced by the assigned conditions to the study. Participant data were analyzed only if all items in the relevant scale were completed. Therefore, minor fluctuations in degrees of freedom for particular analyses exist.

Ratings of the Male Conversant (i.e., Jim)

We predicted that ratings of Jim would depend on all three variables examined in this study. Testing all three hypotheses proposed involved a hierarchical regression analysis with follow-up tests involving participants' judgments of Jim as a function of comment type, instruction set, and bias. ASI scores were mean centered; comment type and instruction set were dummy coded.

As can be seen in Table 2, analysis revealed a significant three-way interaction ($\beta = .41$, $p = .022$, $R^2 = .34$), $F(7, 114) = 8.44$, $p = .001$, $\eta^2 = .05$. When the comment was clearly negative, we predicted that neither gender bias nor correction instructions would be relied upon for judgment. The blatancy of the comment would drive participants' unfavorable ratings of the male conversant. Following statistical procedures outlined by Aiken and West (1991), analysis of the simple slopes comprising the interaction in the negative condition revealed support for Hypothesis 1. Nonsignificant effects for the influence of gender bias in the general instruction set ($\beta = -.08$), $t(114) = -.26$, $SE = .43$, $p = .796$, and correction instruction set ($\beta = .04$), $t(114) = .29$, $SE = .19$, $p = .772$, were obtained.

TABLE 2
Model Summary of Hierarchical Regression Analysis Testing the Influence of Comment Type, Instruction Set, and Gender Bias on Ratings of Male Target

Variable	*B*	*SE*	*β*	$R^2\Delta$
Step 1				.28**
Comment type	–1.05**	.162	–.52	
Instruction set	–.05	.157	–.03	
Gender bias		.03	.108	.02
Step 2				.03
Comment × Instruction		.64*	.322	–.29
Comment × Bias	.01	.221	.01	
Instruction × Bias	–.31	.219	–.34	
Step 3				.031**
Comment × Instruction × Bias	1.03*	.44	.41	

Note. Regression coefficients are reported from the step on which each variable was first entered.
*$p < .05$. **$p < .01$.

Analysis in the ambiguous condition revealed a different pattern. Recall that for Hypothesis 2, we predicted that when the comment was ambiguous, participants would rely on their gender bias to interpret whether he was helpful or condescending. Recall that higher scores on the ASI reflected greater traditional gender bias. Participants with a traditional bias would rate Jim more favorably than participants with an egalitarian gender bias. Results were in support of this hypothesis; in the general instruction set, gender bias positively predicted ratings ($\beta = .97$), $t(114) = 2.66$, $SE = .50$, $p = .009$. Last, we predicted the differences in favorability ratings would be diminished or reversed when instructed to correct for bias (Hypothesis 3). In support of this hypothesis, results revealed that in the correction instruction set, the significant relationship between gender bias and favorability ratings was negative ($\beta = -.42$), $t(114) = -2.01$, $SE = .29$, $p = .047$. Ratings were plotted as a function of comment type and instruction set 1 standard deviation above and below mean ASI scores. In Figure 1, reliance on one's gender bias in the ambiguous condition to guide ratings can be easily seen in the general instruction set. Ratings of Jim's favorability increased in the direction of traditional bias. In the correction instruction set, the direction and magnitude of ratings reversed, demonstrating a correction effect and confirming our third hypothesis. We further tested differences in ratings by instruction set at different levels of gender bias (Jaccard, Turrisi, & Wan, 1990). At 1 standard deviation above the mean score

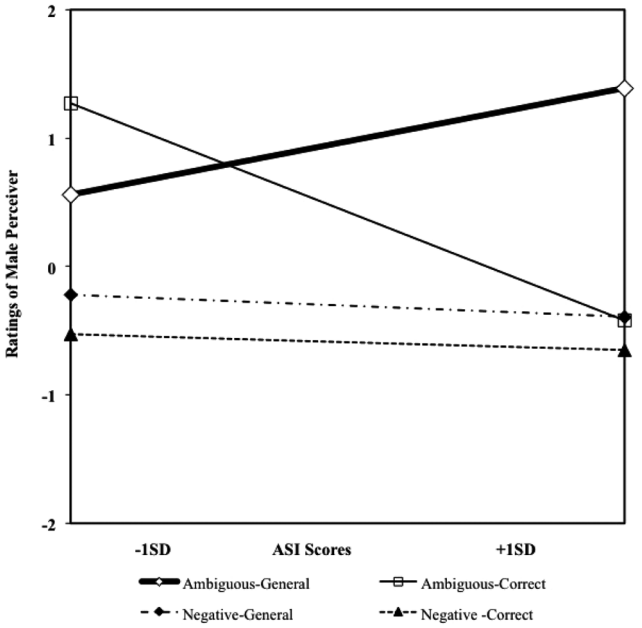

FIGURE 1 Ratings of male perceiver as a function of comment type, instruction set, and Ambivalent Sexism Inventory scores.

for ASI, significant differences in ratings were obtained, $t(49) = 2.43$, $p = .025$. In other words, for participants with a traditional gender bias, their ratings of the male target significantly differed by instruction set. For participants scoring below 1 standard deviation (egalitarian gender bias), the difference in ratings by instruction set was not significant, $t(49) = -.76$, $p = .452$. Therefore, participants successfully adjusted their ratings in the correction condition for the direction of their bias, but the magnitude of their correction was not equivalent for both directions. The adjustment in ratings resulted in a clear overcorrection effect for participants with a traditional gender bias. Participants with an egalitarian bias also showed a correction effect, but the magnitude was not as large. Reasons for this asymmetry are discussed later in the article; however, the primary finding remains: The significant interaction obtained indicates that across levels of bias, participants' ratings of the male target were influenced by the evaluative information about the target and whether they were instructed to correct for their biases or not.

Several aspects of this finding are noteworthy: Even a subtle cue such as names of the conversants was enough to reveal an influence of gender bias; judgments were corrected for in both attitudinal directions when correction instructions were provided. In the negative condition, neither correction instructions nor gender bias influenced judgments in either direction or magnitude. Taken together, these results show that gender biases are more likely to be reflected in judgments and are more susceptible to correction effects when social information is ambiguous and when instructions to correct for bias are provided.

Ratings of Female Conversant (i.e., Ann)

Regardless of instruction set, ratings were similar across bias. Furthermore, neither bias nor experimental conditions influenced participants' ratings of the female conversant. Hierarchical regression analysis using comment, instruction set, and bias showed no significant main effects, $F(1, 121) = 1.03$, $p = .379$, or two-way effects, $F(6, 118) = .57$, $p = .755$, or a three-way effect, $F(7, 117) = .50$, $p = .830$. As designed, participants did not respond differently to her by experimental condition or by gender bias.

DISCUSSION

Past studies have shown differences in social judgments when people were warned that their biases might influence their judgments. The present study examined the influence of instructions to correct for bias when participants were provided with ambiguous or blatant social information. The data provide support for the hypothesis that when given correction instructions that prompt people to think about bias, female participants corrected in the direction of their bias about a male target person but only when the information provided was ambiguous as opposed to blatantly negative. The nonsignificant results obtained in the negative condition affirm that participants determine whether their bias might influence their ratings based on the kinds of social information provided about the target person. Just as our reliance on biases and other subjective factors for social judgments is greater when objective information is unclear or limited, we are also more likely to revise or correct for our biases in these circumstances.

Previous studies have shown bidirectional correction effects within priming paradigms in which participants judged an outgroup member based on their categorical membership. Our study demonstrated that bidirectional correction effects can extend to judgments that rely upon people's awareness of the direction and extent of their biases. In this case, the correction instructions prompted participants to think about whether bias might influence their judgment. Not only did the present study provide more information about the target person than previous studies, but the socially ambiguous information allowed for the application of one's respective bias when asked to judge the target's behavior. The differential impact of the correction instructions depending on the type of social information presented provides additional support for the idea that participants rely upon their own naïve theory to determine if adjustments or corrections are needed. The kinds of judgments and

corrections participants were asked to make in the present study emulate the kinds of scenarios people face in their everyday lives: social interactions with others that require us to infer their intent.

Future Studies on Bidirectional Correction Effects

People correct when they are concerned about being biased or appearing biased to others (Dunton & Fazio, 1997; Olson & Fazio, 2004; Plant & Devine, 1998). One reason for the asymmetry obtained for bidirectional correction effects may be related to perceptions of normative attitudes. Given that egalitarian attitudes are not normatively construed as a bias, it is possible that participants accounted for whether their stance was relatively extreme or socially acceptable before engaging in correction. Additional research is needed to determine if merely warning people about bias engages the correction process or if it occurs only when people believe that their attitude is out of synch with what they perceive is normative. For example, perceived normative sexism moderates men and women's personal level of benevolent sexism (Sibley et al., 2009). A follow-up study manipulating or measuring the perception of normative responses to the male target would provide insight into whether participants with an egalitarian bias believe that their initial judgment is normative, unlike participants with a traditional gender bias who may see their judgments as not normative and therefore are more likely to correct for their bias.

Another promising line of research is to examine the role of social identity and its influence on correction of egalitarian beliefs. Zucker (2004) found distinct differences in the relationship between beliefs and behavior among women who reported egalitarian beliefs but rejected the feminist self-label compared to women who identified as feminist. Given that people with salient social identities are motivated to behave in a manner consistent with their identity (Deaux, Reid, Mizrahi, & Cotting, 1999; Stryker & Serpe, 1994), women who self-identify as feminist may resist instructions to correct compared to women who have these beliefs but do not identify with this label. Future studies delineating differences in personal commitment to egalitarian beliefs might reveal response differences to instructions to correct for bias.

Further study of bidirectional correction effects might also help us understand the reluctance involved in labeling instances of subtle remarks as prejudiced or discriminatory. In a misguided effort to be unbiased, people may engage in a correction process that need not take place. Further experimentation in refinements of instructions is needed to understand the perceiver's perspective on what is meant by bias. One lay definition for bias is an unfair judgment. An alternative definition for bias is simply a tendency or preference, which does not necessarily connote unfairness (Hahn & Harris, 2014). It is possible that, for some people, the term "bias" is strongly associated with unfair outcomes or judgments. Therefore, warnings about bias lead to a perception that any bias, regardless of its valence, is a mental contaminant in need of correction. Hence, advocating the use of correction instructions to produce less prejudiced judgments might inadvertently also lead to the correction of egalitarian views as well. Future studies are needed to test whether different types of correction instructions, such as ones that omit the word "bias," produce the same effects. Researchers can then provide more refined recommendations for how to successfully instruct people to "debias" their judgments in an appropriate manner.

Limitations

Although our primary hypotheses were supported by the data, our evidence for bidirectional correction effects were somewhat weakened by the nonsignificant differences between ratings of the male target for participants with an egalitarian bias. As can be seen in Figure 1 and verified by simple effects tests, the correction instructions had a stronger impact for participants with traditional gender bias compared to participants with an egalitarian bias. Although these rating differences were not significant at these particular points in the data, recall that the interaction and slopes were significant across strength of gender bias. That is, participants adjusted their ratings of the male target according to the direction of their gender bias but did not correct for the same degree across type of bias. Studies exploring why bidirectional effects were asymmetrical are proposed in the future studies section of this article.

Conclusion

Just as stereotype threats exist "in the air" (Steele, 1997), biases exert their influence on social dynamics. Conscious and unconscious motivations and beliefs influence both what we say and what we perceive. The ability to mentally correct our initial judgments opens up the possibility for perceivers to cue into the variety of subtexts that exist in their thoughts and deeds when interacting with others. Ideally, correction processes take place so that egalitarian ideals are promoted, not hindered, but it is only when we recognize and document the power of these dynamics that we can calibrate our mental corrections to ameliorate intergroup interactions.

The primary results advance our theoretical understanding of the correction process by demonstrating that correction instructions that do not explicitly name the bias can lead to an awareness of our biases depending on whether the social context is ambiguous or clear. Previous

research has shown that subjective factors are more likely to exert their influence on our judgments in ambiguous than in clear circumstances. Although it is now apparent that instructions encouraging participants to detect whether their bias might influence their judgments is effective in ambiguous contexts, further studies are needed to produce instructions that encourage correcting for harmful stereotypes but at the same time do not undermine egalitarian ideals.

ACKNOWLEDGMENTS

We thank Marisa Knight and the reviewers for their valuable comments on an earlier version of this article.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Altemeyer, B. (1981). *Right-wing authoritarianism*. Winnipeg, Canada: University of Manitoba Press.

Borgida, E., & Howard-Pitney, B. (1983). Personal involvement and the robustness of perceptual salience effects. *Journal of Personality & Social Psychology*, *45*, 560–570.

Deaux, K., Reid, A., Mizrahi, K., & Cotting, D. (1999). Connecting the person to the social: The functions of social identification. In T. R. Tyler, R. M. Kramer, & O. P. John (Eds.), *The psychology of the social self* (pp. 91–113). Mahwah, NJ: Erlbaum.

DeSteno, D., Petty, R., Rucker, D., Wegener, D., & Braverman, J. (2004). Discrete emotions and persuasion: The role of emotion-induced expectancies. *Journal of Personality and Social Psychology*, *86*, 43–56.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*, 316–326.

Fazio, R. H., & Towles-Schwen, T. (1999). *The MODE model of attitude-behavior processes*. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97–116). New York, NY: Guilford.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*, 77–83.

Fiske, S. T., Lin, M. H., & 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, NY: Guilford.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–74). New York, NY: Academic Press.

Fleming, M. A., Wegener, D. T., & Petty, R. E. (1999). Procedural and legal motivations to correct for perceived judicial biases. *Journal of Experimental Social Psychology*, *35*, 186–203.

Glick, P. (2006). Ambivalent sexism, power distance, and gender inequality across cultures. In *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 283–302). New York, NY: Cambridge University Press.

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. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, *56*, 109–118.

Glick, P., & Fiske, S. T. (2011). Ambivalent sexism revisited. *Psychology of Women Quarterly*, *35*, 530–535.

Glick, P., & Fiske, S. T. (2012). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. In J. Dixon & M. Levine (Eds.), *Beyond prejudice: Extending the social psychology of conflict, inequality and social change* (pp. 70–88). New York, NY: Cambridge University Press.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J., Abrams, D., Masser, B., & Lopez, W. L. (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., ... Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, *86*, 713–728.

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Rationality and motivated reasoning. *The Psychology of Learning and Motivation*, *61*, 41–102.

Higgins, E., Rholes, W., & Jones, C. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141–154.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). Implications of behavioral decision theory and social marketing for designing social action programs. In J. Edwards, R. Tindale, L. Heath, & E. J. Posavac (Eds.). *Social influence processes and prevention* (pp. 103–142). New York, NY: Plenum.

Khan, S., & Lambert, A. (1998). Ingroup favoritism versus black sheep effects in observations of informal conversations. *Basic and Applied Social Psychology*, *20*, 263–269.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution. In E. Higgins, R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior, Vol. 2* (pp. 333–368). New York, NY: Guilford.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theory of stereotype activation and application. *Psychological Bulletin*, *129*, 522–544.

Lambert, A., Khan, S., Lickel, B., & Fricke, K. (1997). Mood and the correction of positive versus negative stereotypes. *Journal of Personality and Social Psychology*, *72*, 1002–1016.

Lepore, L., & Brown, R. (2002). The role of awareness: Divergent automatic stereotype activation and implicit judgment correction. *Social Cognition*, *20*, 321–351.

Marques, J. M., Robalo, E. M., & Rocha, S. A. (1992). Ingroup bias and the 'black sheep' effect: Assessing the impact of social identification and perceived variability on group judgements. *European Journal of Social Psychology*, *22*, 331–352. doi: 10.1002/ejsp.2420220403

Marques, J. M., & Yzerbyt, V. Y. (1988). The black sheep effect: Judgmental extremity towards ingroup members in inter- and intra-group situations. *European Journal of Social Psychology*, *18*, 287–292. doi: 10.1002/ejsp.2420180308

Martin, L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality and Social Psychology*, *51*, 493–504.

Monteith, M. J., Deneen, N. E., & Tooman, G. D. (1996). The effect of social norm activation on the expression of opinions concerning gay men and Blacks. *Basic and Applied Social Psychology*, *18*, 267–288.

Olson, M. A., & Fazio, R. H. (2004). Trait inferences as a function of automatically-activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology*, *26*, 1–11.

Petty, R. E., Priester, J., & Wegener, D. (1994). Cognitive processes in attitude change. In *Handbook of social cognition Volume 2: Applications* (2nd ed., pp. 69–142). Hillsdale, NJ: Erlbaum.

Petty, R. E., & Wegener, D. T. (1993). Flexible correction processes in social judgment: Correcting for context-induced contrast. *Journal of Experimental Social Psychology*, *29*, 137–165.

Petty, R. E., & Wegener, D. T. (1997). Flexible correction processes in social judgment: Implications for persuasion. *Social Cognition*, *16*, 93–113. doi: 10.1521/soco.1998.16.1.93

Plant, E., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*, 741–763. doi: 10.1037/0022–3514.67.4.741

Robins, R., Fraley, R., & Krueger, R. (2007). *Handbook of research methods in personality psychology*. New York, NY: Guilford.

Roets, A., Van Hiel, A., & Dhont, K. (2012). Is sexism a gender issue? A motivated social cognition perspective on men's and women's sexist attitudes toward own and other gender. *European Journal of Personality*, *26*, 350–359. doi: 10.1002/per.843

Schuller, R. A., Kazoleas, V., & Kawakami, K. (2009). The impact of prejudice screening procedures on racial bias in the courtroom. *Law and Human Behavior*, *33*, 320–328. doi: 10.1007/s10979–008-9153–9

Sczesny, S., & Kühnen, U. (2004). Meta-cognition about biological sex and gender-stereotypic physical appearance: Consequences for the assessment of leadership competence. *Personality and Social Psychology Bulletin*, *30*, 13–21.

Sibley, C. G., Overall, N. C., Duckitt, J., Perry, R., Milfont, T. L., Khan, S. S., & Robertson, A. (2009). Your sexism predicts my sexism: Perceptions of men's (but not women's) sexism affects one's own sexism over time. *Sex Roles*, *60*, 682–693.

Sibley, C. G., Wilson, M. S., & Duckitt, J. (2007). Antecedents of men's hostile and benevolent sexism: The dual roles of social dominance orientation and right-wing authoritarianism. *Personality And Social Psychology Bulletin*, *33*, 160–172. doi: 10.1177/0146167206294745

Sommers, S. R. (2006). On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology*, *90*, 597–612. doi: 10.1037/0022–3514.90.4.597

Stapel, D. A., Martin, L. L., & Schwarz, N. (1998). The smell of bias: What instigates correction processes in social judgments? *Personality and Social Psychology Bulletin*, *24*, 797–806. doi: 10.1177/0146167298248002

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Strack, F. (1992). The different routes to social judgments: Experimental versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 249–276). Hillsdale, NJ: Erlbaum.

Strack, F., & Mussweiler, T. (2001). Resisting influence: Judgmental correction and its goals. In J. P. Forgas & K. D. Williams (Eds.), *Social influence: Direct and indirect processes* (pp. 199–212). New York, NY: Psychology Press.

Stryker, S., & Serpe, R. T. (1994). Identity salience and psychological centrality: Equivalent, overlapping, or complementary concepts? *Social Psychology Quarterly*, *57*, 16–35.

Swim, J. K., Scott, E. D., Sechrist, G. B., Campbell, B., & Stangor, C. (2003). The role of intent and harm in judgments of prejudice and discrimination. *Journal of Personality and Social Psychology*, *84*, 944–959.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, *68*, 36–51.

Wegener, D. T, & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. In *Advances in experimental social psychology, Vol. 29* (pp. 141–208). San Diego, CA: Academic Press.

Wilson, T., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*, 117–142.

Wilson, T., Centerbar, D., & Brekke, N. (2002). Mental contamination and the debiasing problem. *Heuristics and biases: The psychology of intuitive judgment* (pp. 185–200). New York, NY: Cambridge University Press.

Wyer, R. S., & Budesheim, T. L. (1987). Person memory and judgments: The impact of information that one is told to disregard. *Journal of Personality and Social Psychology*, *53*, 14–29.

Zucker, A. N. (2004). Disavowing social identities: What it means when women say, "I'm not a feminist, but..." *Psychology of Women Quarterly*, *28*, 423–435.