


Exaggerating Accessible Differences: When Gender Stereotypes Overestimate Actual Group Differences

Personality and Social
Psychology Bulletin
2017, Vol. 43(9) 1323–1336
© 2017 by the Society for Personality
and Social Psychology, Inc
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167217713190
journals.sagepub.com/home/pspb


Tal Eyal¹ and Nicholas Epley²

Abstract

Stereotypes are often presumed to exaggerate group differences, but empirical evidence is mixed. We suggest exaggeration is moderated by the accessibility of specific stereotype content. In particular, because the most accessible stereotype contents are attributes perceived to differ between groups, those attributes are most likely to exaggerate actual group differences due to regression to the mean. We tested this hypothesis using a highly accessible gender stereotype: that women are more socially sensitive than men. We confirmed that the most accessible stereotype content involves attributes perceived to differ between groups (pretest), and that these stereotypes contain some accuracy but significantly exaggerate actual gender differences (Experiment 1). We observe less exaggeration when judging less accessible stereotype content (Experiment 2), or when judging individual men and women (Experiment 3). Considering the accessibility of specific stereotype content may explain when stereotypes exaggerate actual group differences and when they do not.

Keywords

stereotypes, stereotype accuracy, gender differences, social sensitivity, social judgment

Received July 29, 2016; revision accepted May 9, 2017

Few beliefs are more frequently maligned by psychologists and laypeople alike than stereotypes. Faulty (Allport, 1954), biased (Fiske & Taylor, 1984), and hateful (Stangor & Schaller, 1996) are just a few of the published depictions. When stereotypes contain some degree of accuracy, they are still widely presumed to exaggerate the magnitude of actual group differences (Ford & Stangor, 1992).

Even when there is a “kernel” of truth to a stereotype, stereotypes are typically stronger and more pervasive than the kernel would justify . . . , presumably because the strength and consistency of a phenomenon are exaggerated in perceivers’ minds. (Hall, Coats, & LeBeau, 2005, p. 914)

Intergroup evaluations thus seem guided more by figments of imagination than by facts of life.

In contrast to the common perception that stereotypes exaggerate group differences, empirical tests of whether stereotypes *actually* exaggerate group differences yield mixed results. These tests typically compare predictions of group differences against actual group differences, some of which show evidence of exaggeration (Beyer, 1999; Chambers & Melnyk, 2006; Corneille & Judd, 1999; Dawes, Singer, & Lemons, 1972; Graham, Nosek, & Haidt, 2012; Krueger, Hasman, Acevedo, & Villano, 2003; Martin, 1987; Quellar, Schell, & Mason, 2006), whereas others do not (Ashton &

Esses, 1999; Diekmann, Eagly, & Kulesa, 2002; McCauley, Thangavelu, & Rozin, 1988). After a comprehensive review of the empirical evidence, Jussim (2012) concluded,

The strong form of the exaggeration hypothesis—either defining stereotypes as exaggerations or claiming that stereotypes lead to exaggeration—is dead. Exaggeration does sometimes occur, but it does not appear to occur much more frequently than accuracy or underestimation, and may even occur less frequently. (p. 392)

Based on this evidence, stereotypes appear to be reasonably calibrated, with mistakes that underestimate and overestimate group differences arising more or less randomly. Intergroup perceptions seem to be guided by the facts of life, plus some random errors of imagination.

Here, we suggest a potential resolution to these two seemingly opposing viewpoints by testing a theory about when a group stereotype will exaggerate the differences between groups and when it will not. Research on stereotype accuracy

¹Ben Gurion University of the Negev, Beer-Sheva, Israel

²University of Chicago, IL, USA

Corresponding Author:

Tal Eyal, Department of Psychology, Ben Gurion University of the Negev, Building 98, Beer-Sheva, 84105, Israel.
Email: taleyal@bgu.ac.il

has long been disconnected from theories about the mechanisms that guide human judgment, often yielding a collection of findings that are tallied rather than an understanding about when social judgments are likely to be accurate and when they are not (Gilbert, 1998). Understanding the mechanisms that guide judgment should enable a more precise understanding of the relative accuracy and inaccuracy of these judgments (Eyal & Epley, 2010; Funder, 1995; Vazire, 2010).

In particular, we predict that stereotype exaggeration is not haphazard and unpredictable, but rather that some stereotype content will reliably exaggerate group differences whereas other stereotype content will not. Specifically, we predict that the most accessible stereotype content—the content that typically defines a given group’s differentiating attributes—is likely to systematically exaggerate actual group differences but that less accessible content is not. Those who malign stereotypes for exaggerating group differences often focus on the most accessible components of a stereotype, whereas researchers who study accuracy often study a much broader range of attributes that vary in their accessibility. Both positions could be right because they are evaluating different stereotype content.

Our prediction applies only to one particular kind of stereotype inaccuracy: exaggerating the actual difference between two groups on a given attribute. This differs from the type of stereotype exaggeration typically measured by researchers. Judd and Park (1993), for instance, identified three other forms of stereotype inaccuracy that apply to evaluations of a single group: misplacing the level of a given group attribute in a direction consistent with a stereotype (i.e., stereotype inaccuracy), mistaking the valence of a group attribute (i.e., valence inaccuracy), or underestimating the variance on an attribute within a group (i.e., dispersion inaccuracy or outgroup homogeneity). Each of these compares evaluations of a target group against a specific outcome measure, whereas our analysis compares the *difference* in predicted performance of two groups against the actual *difference* between those two groups.

Our theory predicts that exaggerating the differences between groups is not random but rather systematic. Again, those who malign stereotypes for exaggerating group differences could be referring to a more systematic form of inaccuracy, whereas researchers who study it may be measuring other forms. Precision about both stereotype content and the particular form of exaggeration it implies should yield a more precise understanding of the factors that moderate stereotype exaggeration. Here, we report a series of experiments testing our hypotheses using gender stereotypes.

Different, by Definition

By definition, a stereotype is simply “a set of beliefs about the personal attributes of a group of people” (Ashmore & Del Boca, 1981, p. 16). Technically speaking, a stereotype’s content could then include beliefs about any possible attribute one might apply to any group, from the degree to which a

group prefers apples over oranges to the degree to which a group prefers liberal over conservative policies. Practically speaking, some stereotype content comes to mind more easily and is more frequently associated with a group than other content. When Americans think of Republicans and Democrats, for instance, the first beliefs that come to mind are political preferences rather than fruit preferences.

This occurs because groups are defined by the attributes that make them different from other groups (Judd & Park, 1993), just as the self is defined by attributes that make an individual different from other individuals (McGuire & McGuire, 1988; Mussweiler, 2002; Nelson & Miller, 1995). Republicans are conservative whereas Democrats are liberal. Baptists believe in God but atheists do not. Pro-life advocates oppose abortion rights whereas pro-choice advocates defend them. Women are more socially sensitive but less physically aggressive than men. Defining a group by the attributes that make it *similar* to other groups—such as its members prefer autonomy over tyranny, would rather be in love than in war, or care deeply for their own children—means that you have defined the group out of existence. As a result, the most accessible content of a given stereotype will include attributes that are presumed to differentiate one group from another rather than attributes that both groups are presumed to share. When people spontaneously think of men, for instance, they are likely to think of attributes that make men *different* from women (e.g., aggression, socially insensitivity) rather than of attributes presumed to be roughly similar (e.g., impulsiveness, general happiness).

If groups are defined by their differences rather than by their similarities, then the most highly accessible stereotype content should systematically exaggerate actual group differences, even if those stereotypes are moderately accurate. Indeed, existing research makes it clear that beliefs about the attributes of groups are positively, but imperfectly, correlated with actual group attributes (Diekmann et al., 2002; Judd, Ryan, & Park, 1991; Martin, 1987; Wolsko, Park, Judd, & Wittenbrink, 2000). Whenever two variables are imperfectly correlated, high scores on one variable are associated with more moderate scores on the other variable by statistical necessity. Galton was the first to notice this “result of timeless significance” (Edwards, 1993, p. 93) when observing that tall fathers tended to have tall sons, but the tallest fathers tended to have slightly shorter sons. Likewise, short fathers tended to have short sons, sons who were slightly taller than they were. Although every student of statistics now learns of “regression to the mean,” recognizing its operation in everyday life, as well as in science, is notoriously difficult (Harrison & Bazerman, 1995; Krueger, Savitsky, & Gilovich, 1999).

Overview of Current Research

We believe the psychological processes that create stereotype content can help to explain when a specific belief about a group will exaggerate actual group differences and when it

will not. To the extent that stereotypes contain some accuracy about real group differences in the world (Jussim, Crawford, & Rubenstein, 2015), and the attributes people tend to associate with a group are those that differentiate one from another, then the most readily accessible attributes of a group should also be most likely to exaggerate group differences. Less accessible attributes should be less likely to exaggerate group differences.

We tested this hypothesis in five experiments that measured what is presumed to be one of the largest psychological gender differences: that women are more socially sensitive than men (more empathic, and better able to understand others' thoughts and feelings). Indeed, Baron-Cohen (2003) described this as the "essential difference" between men and women.

Whereas previous research on gender stereotypes' accuracy has mainly compared people's estimations of men's and women's behaviors with self-report measures from separate surveys (e.g., Briton & Hall, 1995; Diekmann et al., 2002; Wolsko et al., 2000), the current research compares people's evaluations of men's and women's performance on specific tests with their actual performance on those same tests. This therefore avoids the major potential problem of biased self-reporting of actual behavior (Judd & Park, 1993). More important, we go beyond prior work by attempting to explain when beliefs about gender differences will be relatively calibrated and when they will exaggerate differences between groups. Because social sensitivity is perceived to strongly differentiate men from women, we predicted it would be among the most highly accessible components of gender stereotypes, and would also significantly exaggerate actual group differences. We predicted that such exaggeration would be more pronounced in evaluations of groups than in evaluations of individual men and women. We also predicted that exaggeration would be reduced among less accessible components of gender stereotypes, because this stereotype content is not presumed to differ between men and women.

Pretest

We conducted a pretest to confirm our prediction that groups are defined by their perceived differences rather than by perceived similarities (Judd & Park, 1993), that these perceived differences are the most accessible components of gender stereotypes, and that social sensitivity is presumed to be a large gender difference.

Two hundred four MTurk workers participated in an online experiment in exchange for a small fee (102 women, $M_{\text{age}} = 35.53$, $SD_{\text{age}} = 12.44$). Participants read a list of 30 attributes (15 pairs of antonyms) taken mostly from Hyde (2005), presented in a random order. Six attributes were related to social sensitivity (high social sensitivity/low social sensitivity, empathic/individualistic, compassionate/inconsiderate) whereas the remaining attributes were not (e.g., low math ability/high math ability). To measure stereotype accessibility, participants in the category condition went through

the list and selected all adjectives that were part of the common stereotype of men (men condition) or women (women condition). Those that are considered to be part of a stereotype would be those that come to mind when thinking about men and women. Participants then saw only those adjectives they selected and rank ordered them from the adjective they thought is most strongly believed to be typical of men/women (i.e., most accessible for the stereotype of men/women) to the adjective from the list that is least strongly believed to be typical of men/women. Participants then rated how descriptive each attribute was of men in general and women in general (1 = *not at all*, 2 = *somewhat*, and 3 = *very much*). For half of the participants, the order of tasks was different—first they rated the descriptiveness of each attribute, then selected the attributes that were part of the stereotype, and then ranked the selected attributes. We also randomly varied the order in which participants evaluated each gender (some evaluated men first, others women first). Finally, all participants reported the perceived difference between men and women on each attribute (1 = "men are more so than women," 2 = "men and women are the same," and 3 = "women are more so than men").

To further examine stereotype calibration, we also included an individual condition. Participants in this condition were asked to think about a specific man or woman they knew well (e.g., a friend, a relative, a colleague), write the initials of that man (e.g., MM) or woman (e.g., FF), select all adjectives that are part of the description of the specific man or woman, rank order them from the most to the least descriptive attribute, and then rate how descriptive each attribute is of that man or woman (1 = *not at all*, 2 = *somewhat*, and 3 = *very much*). Half of the participants went through this procedure in a reversed order—they first rated the descriptiveness of each attribute, then selected and rank ordered the most descriptive attributes of the individual. Finally, all participants reported the perceived difference between the specific man and woman on each attribute (1 = "MM is more so than FF," 2 = "MM and FF are the same," and 3 = "FF is more so than MM"). Because stereotypes are beliefs about groups of people rather than about known individuals of a group, we expected larger predicted gender differences on attributes related to social sensitivity in the category condition than in the individual condition.

The number of selected attributes varied widely across participants and so we first identified the five attributes each participant ranked highest and then reverse scored them so that the top ranked attribute was scored as a 5 and the lowest ranked attribute as a 1. To test whether accessible stereotype content is based on perceived group differences, we calculated the difference in the predicted mean ratings and rankings of descriptiveness for men and women (category condition) for each attribute, and also for a man and a woman (individual condition). As predicted, for the difference in predicted mean ratings of descriptiveness, the measure of stereotype content was almost perfectly correlated with the size of perceived gender differences in the category

condition, $r(30) = .99$, but these measures were significantly less correlated in the individual comparisons, $r(30) = .65$, $z = 6.88$, $p < .001$. For the difference in predicted mean rankings of descriptiveness, the stereotype content was more strongly correlated in the category condition, $r(30) = .83$, than in the individual comparisons, $r(30) = .66$, but this difference was nonsignificant, $z = 1.45$, $p = .146$. Overall, the attributes presumed to differ between men and women are the defining features of men and women.

We expected larger perceived gender differences in the category condition than in the individual condition. To test this, we calculated the percentage of participants who selected each attribute as descriptive of men/women and of a man/woman (see Supplemental Figure 1). We then calculated the absolute difference between the percentage of participants who selected each attribute as descriptive of men versus women and of a man versus woman. As expected, we observed significantly larger gender differences in the category condition ($M = 37.77$) than in the individual condition ($M = 15.90$), $t(58) = 4.58$, $p < .001$, $r = .51$. In addition, the most pronounced differences were obtained for attributes related to social sensitivity (i.e., have high social sensitivity, empathic, compassionate, and warm). These results suggest that particularly large perceived differences are a function of group stereotypes.

We also examined whether the predicted differences between the rankings of attributes are larger when evaluating gender categories than when evaluating individuals. We then calculated the absolute difference between the rankings of each trait. As expected, participants predicted larger gender differences when evaluating categories ($M = 0.80$) than when evaluating individuals ($M = 0.17$), $t(58) = 4.92$, $p < .001$, $r = .53$ (see Supplemental Figure 2). In addition, the largest differences in ranking of attributes of men and women again involved social sensitivity. We also examined whether the predicted differences between the ratings of attributes are larger when evaluating gender categories than when evaluating individuals. We thus calculated the absolute difference between the ratings of traits descriptive of women and men and of a woman and a man. Perceived gender differences were larger for the category ($M = 0.68$) than for the individual ($M = 0.22$), $t(58) = 6.74$, $p < .001$, $r = .66$ (see Supplemental Figure 3).

Finally, we examined whether attributes related to social sensitivity yielded the largest perceived group differences (see Supplemental Figure 4). Out of the top five attributes perceived to be more descriptive of women than men, four were related to social sensitivity: have high social sensitivity, compassionate, empathic, and warm. Of the top five attributes perceived to be more descriptive on average of men than women, two attributes were related to social sensitivity: have low social sensitivity and inconsiderate.

The results of the pretest confirm that the most accessible content of gender stereotypes are attributes perceived to differ between men and women. In addition, women are perceived to be more socially sensitive than men, and this is one

of the strongest stereotypes about men and women. We next test the degree to which that stereotype exaggerates actual group differences.

Experiments 1a to 1c

Participants completed one of three widely used social sensitivity tests and then predicted the performance of men and women on that test who were also enrolled in the experiment. These measures avoid several of the most common problems associated with stereotype accuracy research by using a concrete measure of performance, by explicitly restricting predictions to the sample of participants who are actually taking the test, and by using a reasonably representative sample for both predicted and actual performance. It is also important to note that our main predications are not about whether people overestimate or underestimate a given gender's degree of social sensitivity (referred to as "elevation" in the accuracy literature), but rather whether people exaggerate the *difference* between genders on these measures.

This comparison of two difference scores avoids well-known problems in evaluating simple discrepancy scores between a judgment and a criterion measure (judgment-elevation, judgment-positivity, or judgment extremity biases: Cronbach, 1955; Judd & Park, 1993). These problems are especially severe when evaluating a measure that has little intuitive meaning for a participant, such as the number of items correct on a novel test. We therefore do not interpret simple discrepancies between predicted and actual performance for men and women separately, as those simple discrepancy scores are hard to interpret. Instead, we predicted that presumed *differences* between men and women would be significantly larger than the actual *differences* between men and women. However, we present analyses using the Judd and Park (1993) method in the appendix for interested readers.

Method

Participants. MTurk workers participated in exchange for a small fee in Experiment 1a ($N = 233$, 126 women, $M_{\text{age}} = 35.21$, $SD_{\text{age}} = 12.57$), Experiment 1b ($N = 231$, 121 women, $M_{\text{age}} = 34.87$, $SD_{\text{age}} = 12.33$), and Experiment 1c ($N = 238$, 112 women, $M_{\text{age}} = 35.56$, $SD_{\text{age}} = 12.60$). We aimed for large samples recruited from a reasonably representative subject population to obtain reliable measures of predicted and actual gender differences.

Procedure. Participants first read that they would be completing a "social intelligence test" and that social intelligence was important for a "variety of desirable outcomes in daily life . . ." Participants in Experiment 1a then completed the facial expression version of the Diagnostic Analysis of Nonverbal Accuracy Scale (DANVA2; Nowicki & Duke, 1994) which consists of 24 pictures of male and female faces expressing an emotion. For each face, participants indicated

Table 1. Means and Standard Deviations for Experiments 1a to 1c.

Experiment 1a: DANVA2 (24 items)	Women (n = 126)	Men (n = 107)	Difference
Predicted self-performance	17.94 (4.18)	18.04 (3.49)	-0.10 ^a
Predicted others' performance	18.55 (4.12)	15.86 (4.16)	2.69 ^b
Actual performance	19.09 (2.48)	18.67 (3.04)	0.42 ^a
Experiment 1b: Mind in Eyes (36 items)	Women (n = 121)	Men (n = 110)	Difference
Predicted self-performance	23.21 (6.78)	23.22 (6.42)	-0.01 ^a
Predicted others' performance	24.72 (5.89)	20.19 (5.34)	4.53 ^b
Actual performance	26.69 (4.34)	24.36 (6.45)	2.33 ^c
Experiment 1c: Empathy Quotient (88 items)	Women (n = 112)	Men (n = 126)	Difference
Predicted self-performance	64.29 (14.88)	61.36 (13.90)	2.93 ^a
Predicted others' performance	68.71 (11.00)	54.32 (10.18)	14.39 ^b
Actual performance	66.11 (10.16)	64.33 (9.42)	1.78 ^a

Note. For predicted self-performance and actual performance, gender differences are between participants whereas for predicted others' performance, gender differences are within participants. Within each experiment, numbers that do not share a superscript within the difference column differ significantly at $p < .05$.

whether the person feels happy, sad, angry, or fearful. Participants then predicted how many of the 24 items they answered correctly¹ as well as how many items “men MTurk workers participating in this study, on average” and “women MTurk workers participating in this study, on average” answered correctly.²

Participants in Experiment 1b then completed the Reading the Mind in the Eyes test (ME; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). This consists of 36 black and white pictures of the area around the eyes of males and females. Participants chose which of four words (e.g., serious, ashamed, alarmed, bewildered) described what the person in the picture is thinking or feeling. Participants then predicted how many of the 36 items they answered correctly, and how many “men MTurk workers participating in this study, on average” and “women MTurk workers participating in this study, on average” answered correctly.

Participants in Experiment 1c then completed the Empathy Quotient–Short (EQ; Baron-Cohen & Wheelwright, 2004; Wakabayashi et al., 2006) in which participants report their agreement with 22 statements, 16 of which reflect high empathy (e.g., “I find it easy to put myself in someone else’s shoes”) and six which reflect low empathy (e.g., “It is hard for me to see why some things upset people so much”). Participants then predicted their own score on the test and the score for “men MTurk workers participating in this study, on average” and “women MTurk workers participating in this study, on average.”

Note that we restricted predictions to other MTurk workers to enable more appropriate comparisons between predicted and actual performance.

Results and Discussion

Means and standard deviations for the three experiments are presented in Table 1.

In Experiment 1a, we did not observe a significant difference in actual performance on the DANVA2 between women and men ($M_s = 19.09$ and 18.67 , respectively), $t(231) = 1.15$, $p = .253$, 95% confidence interval (CI) $[-.29, 1.13]$, $r = .08$, but participants predicted that women would perform significantly better than men ($M_s = 18.55$ and 15.86 , respectively), $t(232) = 12.57$, $p < .001$, 95% CI $[2.26, 3.10]$, $r = .64$. To test whether predictions exaggerated the actual gender difference, we created a simple calibration score for each participant by subtracting the actual average difference in performance between men and women (0.42) from his or her predicted difference in performance between men and women ($M = 2.69$). This resulting difference score was significantly positive, $t(232) = 10.60$, $p < .001$, 95% CI $[1.84, 2.68]$, $r = .57$, indicating that the predicted gender difference was significantly larger than the actual gender difference.

To test whether this exaggeration effect is robust beyond this one actual group difference score obtained by this particular sample of participants, we employed the bootstrapping method using R to obtain something closer to a population sample. We calculated 1,000 samples with replacement from the original sample, each the same size as the original, and computed a calibration score for each sample. The bootstrap confidence interval (CI) did not include zero $[1.43, 3.09]$, indicating significant exaggeration.

It could be argued that our calibration scores are invalid because they compare a single score of actual performance with a prediction score created by the difference between two scores (i.e., predicted women’s performance and predicted men’s performance). To test whether the exaggeration effect is also robust when actual and predicted differences are treated on more even grounds, we ran a different bootstrapping procedure with 1,000 samples with replacement, the same size as the original, and computed the difference between actual performance of men and women, the difference between

predicted performance between men and women (using only the first prediction that participants gave), and the difference between these two difference scores. The CI was very similar to the other bootstrap analysis we conducted and did not include zero [0.56, 3.19], indicating significant exaggeration.

In Experiment 1b, women answered more items on the ME test correctly than men ($M_s = 26.69$ and 24.36 , respectively), $t(229) = 3.25$, $p = .001$, 95% CI [0.91, 3.74], $r = .21$, but participants predicted an even larger gender difference ($M_s = 24.72$ and 20.19 , respectively), $t(230) = 11.14$, $p < .001$, 95% CI [3.71, 5.33], $r = .59$. We calculated a calibration score for each participant as in Experiment 1a. The resulting score was significantly positive, $t(230) = 5.41$, $p < .001$, 95% CI [1.40, 3.00], $r = .34$, indicating that predicted gender difference on the ME test was significantly larger than the actual gender difference. In addition, the bootstrap CI did not include zero [0.47, 3.91], indicating significant exaggeration. As in Experiment 1a, we also conducted a bootstrapping procedure with the difference between actual performance of men and women, the difference between predicted performance between men and women (using only the first prediction that participants gave), and the difference between these two difference scores. The CI did not include zero [0.44, 4.60], indicating significant exaggeration.

In Experiment 1c, we first calculated each participant's EQ score by summing the 22 items (reverse scoring the six negatively framed items, Cronbach's $\alpha = .93$). We observed no significant difference between the actual EQ scores of women and men ($M_s = 66.11$ and 64.33 , respectively), $t(236) = 1.40$, $p = .162$, 95% CI [-1.27, 1.78], $r = .09$, but participants expected that women would score higher than men ($M_s = 68.71$ and 54.32 , respectively), $t(237) = 16.41$, $p < .001$, 95% CI [12.67, 16.12], $r = .73$. The calibration score was significantly positive, $t(237) = 14.38$, $p < .001$, 95% CI [10.88, 14.34], $r = .68$, indicating that predicted gender differences in EQ was significantly larger than the actual gender difference. In addition, the bootstrap CI did not include zero [9.34, 15.72], indicating significant exaggeration. As in Experiments 1a and 1b, we conducted another bootstrapping procedure with the difference between actual EQ of men and women, the difference between predicted EQ between men and women (using only the first prediction that participants gave), and the difference between these two difference scores. The CI also did not include zero [8.61, 16.01], indicating significant exaggeration.

Women are presumed to be more socially sensitive than men: better able to recognize others' emotions, understand others' intentions, and empathize with others' experiences. Three experiments demonstrate that these gender stereotypes significantly exaggerate actual gender differences between men and women on three relatively objective measures of social sensitivity. In each experiment, we observed statistically nonsignificant differences in the actual performance of men and women on tests of social sensitivity, but large differences in predicted performance. Gender stereotypes were not

directionally mistaken, but they were dramatically miscalibrated. Gender stereotypes suggest that "Men are from Mars and Women are from Venus" (Gray, 1992). The truth is closer to "Men are from Boston and Women are from New York" (see also Hyde, 2005).

Experiment 2

The psychological processes that create stereotypes should help to explain differences in the degree to which they exaggerate actual differences. To the extent that stereotypes contain some degree of accuracy, those attributes perceived to differ most strongly between groups, and hence form the most accessible stereotype content, should exaggerate group differences the most based on regression to the mean alone. We therefore predicted that the most accessible attributes of a stereotype—those that differentiate one group from another most strongly (Judd & Park, 1993)—exaggerate actual group differences but that less accessible content that does not differentiate one group from another will be less likely to exaggerate group differences. People's beliefs about groups therefore do not necessarily exaggerate differences between groups, but some beliefs about groups are more likely to do so than others.

We tested this hypothesis by examining perceived versus actual gender differences on traits that presumed to differentiate men from women according to our pretest (social sensitivity and self-esteem) and one trait presumed to be similar between men and women (happiness). We selected these attributes because each is relatively straightforward to measure with widely accepted and well-validated methods. We expected that participants would exaggerate gender differences when predicting social sensitivity and self-esteem, but not when predicting happiness.

Method

Participants. MTurk workers participated in exchange for a small fee ($N = 132$, 50 women, $M_{age} = 30.98$, $SD_{age} = 10.35$).

Procedure. We first measured participants' stereotypes to confirm that attributes related to social sensitivity and self-esteem, but not to happiness, are perceived to differ between men and women as we found in the pretest described earlier. Participants read the list of 30 attributes used in the pretest and indicated what the stereotype suggests about the similarities or differences between men and women (1 = "stereotype is that men are much more so than women," 5 = "stereotype is that women are much more so than men" with a middle point of 3 = "stereotype is that men and women are the same"). There was also a sixth option that nobody chose ("This attribute is not included in the stereotype").

Participants then read that they will complete several short tests: a facial expressions test that measures empathy (DANVA2), a self-esteem test (Rosenberg, 1965), and a happiness test (Lyubomirsky & Lepper, 1999). The self-esteem

Table 2. Means and Standard Deviations for the Three Tests in Experiment 2.

	Women	Men	Difference
DANVA2 (24 items)			
Predicted self-performance	18.42 (3.23)	16.99 (5.21)	1.43 ^a
Predicted others' performance	18.31 (4.19)	16.08 (4.14)	2.23 ^b
Actual performance	20.02 (1.74)	18.98 (2.58)	1.04 ^a
Self-esteem (10 items, Rosenberg, 1965)			
Predicted self-performance	28.36 (8.10)	29.00 (8.08)	-0.64 ^a
Predicted others' performance	28.14 (5.46)	31.96 (5.10)	-3.82 ^b
Actual performance	19.86 (7.65)	19.26 (6.29)	0.60 ^a
Happiness (4 items, Lyubomirsky & Lepper, 1999)			
Predicted self-performance	19.75 (6.36)	18.90 (6.30)	0.85 ^a
Predicted others' performance	20.08 (4.16)	20.71 (3.50)	-0.63 ^a
Actual performance	17.86 (5.74)	18.80 (5.69)	-0.94 ^a

Note. For predicted self-performance and actual performance, gender differences are between participants whereas for predicted others' performance, gender differences are within participants. Within each test, numbers that do not share a superscript within the difference column differ significantly at $p < .05$.

test is a 10-item measure in which participants report the degree to which they agree with a series of self-descriptive statements on scales ranging from 1 (*strongly disagree*) to 4 (*strongly agree*) ("I feel that I am a person of worth, at least on an equal plane with others"). The happiness test is a four-item measure in which participants respond on scales ranging from 1 (*not a very happy person*) to 7 (*a very happy person*; e.g., "in general I consider myself . . ."). We counter-balanced the order of the tests across participants.

Finally, participants predicted their own performance as well as the performance of women and men, on average, on each of the three tests. For the DANVA2, participants predicted their own total score on the test as well as the average score for men and for women, as done in Experiment 1. For the self-esteem test, participants predicted their score: "On this range from highest on self-esteem being 40 to lowest self-esteem being 10, what do you think was YOUR score?" as well as "the average score for MEN on this test" and "the average score of WOMEN on this test." For the happiness measure, participants predicted their score: "On this range from highest on happiness being 28 to lowest happiness being 4, what do you think was YOUR score?" as well as "the average score for MEN on this test" and "the average score of WOMEN on this test."

Results and Discussion

Means and standard deviations for the three tests are presented in Table 2.

Stereotypes. We examined whether attributes related to social sensitivity, self-esteem, and happiness yielded perceived group differences. As shown in Figure 1, of the 12 attributes where women were expected to show more of the attribute than men, attributes related to social sensitivity were ranked second, third, and sixth: empathic, compassionate, and high social sensitivity. Low self-esteem ranked 11th. Women and

men were not presumed to differ significantly in depression or happiness. Of the 10 attributes where men were expected to show more of the attribute than women, attributes related to social sensitivity were ranked fourth, sixth, and 10th: inconsiderate, low social sensitivity, and individualistic. High self-esteem was ranked eighth. Consistent with the pretest described earlier, men and women were presumed to differ in their attributes related to social sensitivity and self-esteem, but not in their general happiness.

Actual versus predicted performance. Women answered more items on the DANVA2 correctly than men ($M_s = 20.02$ and 18.98 , respectively), $t(130) = 2.53$, $p = .013$, 95% CI [0.23, 1.86], $r = .22$, but participants predicted an even larger gender difference ($M_s = 18.31$ and 16.08 , women vs. men respectively), $t(131) = 5.88$, $p < .001$, 95% CI [1.48, 2.98], $r = .46$. We calculated a calibration score for each participant as in Experiment 1. This score was significantly positive, $t(131) = 3.13$, $p = .002$, 95% CI [0.44, 1.94], $r = .26$, indicating that predicted gender differences in social sensitivity were significantly larger than actual gender differences. In addition, the bootstrap CI did not include zero [0.06, 2.36], indicating significant exaggeration.

To assess self-esteem, we first reverse scored the relevant items and then summed the items into a total score (Cronbach's $\alpha = .94$). There was no significant difference in actual self-esteem between women and men ($M_s = 19.86$ and 19.26 , respectively), $t(130) = -.49$, $p = .623$, 95% CI [-1.82, 3.03], $r = .04$, but participants predicted a significant gender difference ($M_s = 28.14$ and 31.96 , women vs. men respectively), $t(131) = -6.61$, $p < .001$, 95% CI [-4.97, -2.68], $r = .50$. This predicted difference was similar in effect size to the predicted difference on the DANVA2. The calculated calibration score was significantly positive, $t(131) = -7.64$, $p < .001$, 95% CI [-5.57, -3.28], $r = .56$, indicating that predicted gender differences in self-esteem were significantly larger than actual differences. In addition, the bootstrap CI

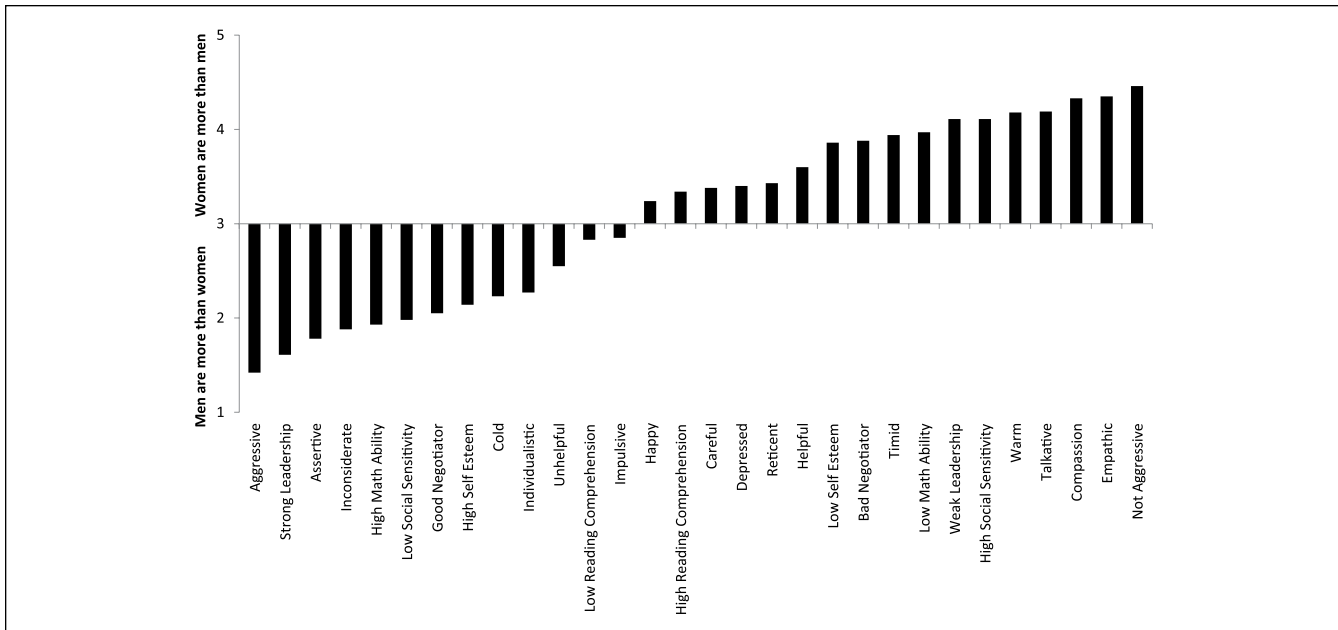


Figure 1. Stereotypes regarding men and women (Experiment 2).

did not include zero $[-7.33, -1.24]$, indicating significant exaggeration.

To assess participants' happiness, we reverse scored the relevant items and then summed the items into a total score (Cronbach's $\alpha = .92$). There was no significant difference in happiness scores between women and men ($M_s = 17.86$ and 18.80 , respectively), $t(130) = -.92$, $p = .358$, 95% CI $[-2.97, 1.08]$, $r = .08$, and participants also did not predict a large gender difference ($M_s = 20.08$ and 20.71 , women vs. men respectively), $t(131) = -1.75$, $p = .082$, 95% CI $[-1.34, 0.80]$, $r = .15$. As expected, the calibration score did not differ from zero, $t(131) = 0.87$, $p = .387$, 95% CI $[-0.40, 1.02]$, $r = .08$, indicating that predicted gender differences in happiness did not exaggerate actual gender differences. In addition, the bootstrap CI included zero $[-1.59, 2.23]$, indicating no exaggeration effect.

As in Experiment 1, we conducted an additional bootstrapping procedure for each of the three measures, with the actual difference between men and women, the predicted difference between men and women (using only the first prediction that participants gave), and the difference between these two difference scores. The CI did not include zero for social sensitivity $[0.70, 3.85]$ or for self-esteem $[-8.01, -1.76]$ but did include zero for happiness $[-0.82, 3.43]$, indicating significant exaggeration for social sensitivity and self-esteem but not for happiness.

Exaggeration as a function of magnitude of perceived differences. The precise constellation of traits that come to mind when thinking about groups of individuals obviously varies somewhat from one person to another. To directly test whether exaggeration is stronger for more accessible stereotype content, we examined how individual differences

in accessibility was related to exaggeration of group differences. We used the data from the first part of the experiment in which we measured stereotype content accessibility by having participants rate the extent to which each of 30 attributes (including social sensitivity and self-esteem) is more stereotypical of men versus women. We first computed a social sensitivity composite by averaging together five out of the six attributes related to social sensitivity: empathic, high social sensitivity, considerate, low social sensitivity (reverse coded), and inconsiderate (reverse coded; $\alpha = .80$). We excluded individualistic because it was statistically unrelated to the other items. Larger numbers on this composite indicate that social sensitivity is a more accessibility component of the gender stereotype. We then computed a self-esteem composite by averaging together low self-esteem and high self-esteem (reverse scored; $r = .24$, $p = .007$).³ Larger numbers on each composite score indicate that a given trait is a more accessible component of the gender stereotype.

We conducted two linear regression analyses—one for predicting exaggeration on the DANVA2 by the social sensitivity accessibility score controlling for the low self-esteem accessibility score, and one for predicting exaggeration on self-esteem by the low self-esteem accessibility score controlling for the social sensitivity accessibility score. The first regression yielded a significant effect for social sensitivity accessibility, $\beta = .38$, $t(129) = 4.34$, $p < .001$, and a nonsignificant effect for low self-esteem accessibility, $\beta = -.09$, $t(129) = -1.00$, $p = .318$. The more that social sensitivity was a uniquely accessible component of the gender stereotype, the more participants exaggerated gender differences on the DANVA2. The second regression yielded a significant effect for low self-esteem accessibility, $\beta = -.37$, $t(129) = -4.30$, p

< .001, and a nonsignificant effect for social sensitivity accessibility, $\beta = -.14$, $t(129) = -1.56$, $p = .122$. The more that low self-esteem was a uniquely accessible component of the gender stereotype, the more participants exaggerated gender differences in self-esteem.

As predicted, participants' gender stereotypes exaggerate actual group differences only on the traits that are defining features of gender stereotypes, namely, traits presumed to differentiate men and women. Gender stereotypes exaggerated actual gender differences in measures of social sensitivity and self-esteem, but did not exaggerate actual differences in happiness. We also found support for our prediction when analyzing individual differences in stereotype content accessibility. Exaggeration of gender differences on a given trait (either social sensitivity or self-esteem) was greater the more participants perceived that trait as an accessible stereotype content. Consistent with a regression-to-the-mean account, the degree to which stereotypes exaggerate group differences is a function of the presumed magnitude of perceived differences rather than on actual gender differences.

Experiment 3

Stereotypes are beliefs about groups of individuals, and groups are defined by attributes that differentiate one group from another. Implicit in this theorizing is that members within a given group are defined by a feature they share. Women share a gender category that differentiates them from men on some attributes. Although individuals within the category of "women" all share the same gender, they obviously vary along many other attributes such as age, ethnicity, and education that are not brought to mind when thinking of the broader gender stereotype. This theorizing suggests the exaggerated gender differences we observed in Experiments 1 and 2 are created because stereotype content is based not only on features presumed to differentiate one group from another but also on the defining feature that unites members within a group. This predicts that individuating group members—thinking about a particular man or woman—will reduce the tendency to exaggerate gender differences because other individuating attributes besides gender are more readily accessible than when considering gender alone.

We tested this prediction in Experiment 3 by asking participants to predict the performance of men and women on a social sensitivity test, and also to predict the performance of individual men and women, shown in a picture. Comparing predictions with actual performance of both men and women as a group, as well as with the individuated men and women, tests the calibration of predicted gender differences.

Method

Participants. Visitors to the Museum of Science and Industry in Chicago volunteered to participate in the baseline condition ($N = 30$, 15 women, $M_{\text{age}} = 43.07$, $SD_{\text{age}} = 13.48$) and subsequently in the evaluation condition ($N = 90$, 43 women,

$M_{\text{age}} = 32.53$, $SD_{\text{age}} = 13.18$). We excluded five additional participants from the baseline condition (one was younger than 18, two were not Native English speakers and did not understand the instructions, and two were interrupted by their children). Including these participants, however, does not meaningfully alter the results.

Procedure. The first 30 participants served as our baseline condition so that we could obtain individual photographs to be used by the next 90 participants in the evaluation condition. Participants in the baseline condition completed the DANVA2 and then predicted their own score on the test as well as the average score for men and for women, as in Experiments 1 and 2. Participants then had their picture taken, from the waist up. These baseline participants completed the same procedure as participants in the evaluation condition, except that they did not make predictions about individuated targets.

Participants in the evaluation condition first completed the DANVA2 and then predicted how many items they answered correctly. Participants then made both categorical predictions about how many items "men on average" and "women on average" would answer correctly, and individuated predictions. For the individuated predictions, participants saw a picture of each participant from the baseline condition, and predicted, for each one, how many items he or she answered correctly. The order of categorical and individuated predictions varied randomly across participants.

Results and Discussion

Means and standard deviations are presented in Table 3.

Performance. We observed no reliable gender difference in performance on the DANVA2 ($M_s = 18.05$ vs. 17.48 , respectively), $t(118) = 1.08$, $p = .282$, 95% CI $[-.47, 1.61]$, $r = .10$.

Predicted performance. In the baseline condition, participants predicted that women would perform significantly better on average than men ($M_s = 17.13$ and 13.97 , respectively), $t(29) = 5.75$, $p < .001$, 95% CI $[2.04, 4.29]$, $r = .73$.

In the evaluation condition, we first averaged participants' individuated predictions of the 15 men and 15 women. A 2 (predicted gender: men vs. women) \times 2 (measure: categorical vs. individuated) repeated-measures ANOVA yielded a main effect for predicted gender, $F(1, 89) = 52.39$, $p < .001$, $r = .61$, qualified by a predicted Gender \times Measure interaction, $F(1, 89) = 16.94$, $p < .001$, $r = .40$. Participants predicted women would perform better than men, but this predicted gender difference was roughly 3 times larger when making categorical judgments ($M_s = 17.08$ vs. 14.23 , respectively), $F(1, 89) = 39.72$, $p < .001$, 95% CI $[1.95, 3.74]$, $r = .56$, than when making individuated judgments ($M_s = 16.43$ vs. 15.36), $F(1, 89) = 32.39$, $p < .001$, 95% CI $[0.70, 1.45]$, $r = .52$.

Table 3. Means and Standard Deviations for Experiment 3.

	Women	Men	Difference
DANVA2 (24 items)			
Baseline and evaluation conditions			
Predicted self-performance	16.12 (4.91)	16.00 (5.25)	0.12 ^a
Actual performance	18.05 (2.25)	17.48 (3.35)	0.57 ^a
Evaluation condition			
Predicted others' performance—Category	17.08 (4.51)	14.23 (4.22)	2.85 ^a
Predicted others' performance—Individuals	16.43 (3.43)	15.36 (3.55)	1.07 ^b
Baseline condition			
Predicted others' performance—Category	17.13 (5.83)	13.97 (5.47)	3.17

Note. For predicted self-performance and actual performance, gender differences are between participants whereas for predicted others' performance, gender differences are within participants. Within each condition, numbers that do not share a superscript within the difference column differ significantly at $p < .05$.

We calculated two calibration scores for each participant—for categorical predictions and for individuated predictions. The calibration score was significantly positive for categorical predictions ($M = 2.85$), $t(89) = 5.04$, $p < .001$, 95% CI [1.38, 3.17], $r = .47$, indicating that predicted gender differences exaggerated actual gender differences observed in the entire sample. The calibration score was also significantly positive, yet significantly smaller, for individuated predictions ($M = 1.07$), $t(89) = 2.66$, $p = .009$, 95% CI [0.13, 0.88], $r = .27$. Beliefs about men and women as groups—stereotypes—exaggerated actual gender differences in social sensitivity more than beliefs about individuated men and women. A paired t test comparing the calibration score for categorical predictions and for individuated predictions yielded a significant effect, $t(89) = 4.12$, $p < .001$, 95% CI [0.92, 2.63], indicating that exaggeration was significantly greater for categorical predictions than for individuated predictions.

We also did a bootstrap analysis. For categorical judgments, the bootstrap CI did not include zero [1.26, 4.15], indicating significant exaggeration. However, for individuated judgments, the bootstrap CI did include zero [−0.14, 2.23], indicating nonsignificant exaggeration.⁴

This experiment, again, highlights how the psychological process underlying a person's beliefs affects the calibration of those beliefs. Evaluating individual men and women, based on their picture, reduced the tendency to exaggerate gender differences in social sensitivity, suggesting that the process that creates stereotypic beliefs about *groups* of men and women leads to exaggerated perceptions of gender differences.

General Discussion

Other people are the most complicated agents that most of us will ever think about, and yet, thinking accurately about others is critical for everyday life. It is therefore not surprising that people's beliefs about others contain a good share of both fact and fiction. People are neither bumbling idiots nor brilliant savants around each other, and a psychological

account of social judgment should not simply categorize judgment as one or the other, or tally up a haphazard collection of existing experiments to create some overall average. Our experiments instead tested a theory that connects the mechanisms that guide interpersonal judgment to the accuracy of those judgments, thereby enabling a more precise account of when stereotypes are likely to exaggerate intergroup differences and when they do not.

Because groups tend to be defined by the features that distinguish one from another, we predicted that the most accessible attributes of a stereotype would tend to exaggerate group differences because they are also the features that people perceive to differ most strongly between groups. People's beliefs contain some degree of accuracy but also some error, and whenever two variables are imperfectly correlated with each other, high scores on one variable tend to be associated with more modest scores on another variable. This suggests that the features that are most accessible when thinking about gender categories will exaggerate actual group differences, but those that are less accessible will be less likely to exaggerate differences.

We tested this account by examining the degree to which people's beliefs about fundamental differences between men and women—their interpersonal abilities, specifically their empathic ability to understand the minds of other—exaggerate actual group differences compared with less accessible features in gender stereotypes, such as happiness. In all five experiments, participants consistently believed that women were on average significantly better mind readers than men. These expected gender differences were not completely mistaken, as women were indeed better in some of our tests than men. However, expected gender differences consistently exaggerated the actual magnitude of gender differences, expecting gender differences that were 2 times larger or more than the actual gender differences. These exaggerated beliefs were mostly apparent in stereotypes of men and women in general, were reduced in beliefs about specific men and women, and did not emerge in men's and women's beliefs about themselves. These exaggerated beliefs were also restricted to attributes presumed to be most accessible in

gender stereotypes, and did not emerge on more peripheral attributes such as happiness.

We think these results help clarify a somewhat mixed research literature, with some experiments finding evidence of exaggeration but other experiments finding no evidence of exaggeration (or even underestimating actual group differences; Jussim, 2012). In existing research, stereotype exaggeration is typically explained as resulting from cognitive biases in social information processing (Beyer, 1999; Corneille & Judd, 1999; Dawes et al., 1972; Krueger et al., 2003; Martin, 1987; McCrae et al., 2013). For example, some have suggested that exaggeration results from people's tendency to pay more attention to and weigh more heavily group members that better differentiate between two social groups (Corneille & Judd, 1999; Eiser, 1971; Tajfel & Wilkes, 1963). Others have suggested that people overestimate the frequency of occurrence of schema-consistent information relative to schema-inconsistent information (Martin, 1987).

We suggest instead that stereotype exaggeration results in part from the way people perceive groups. Any two groups of human beings have more shared attributes than differing attributes, and yet, groups are defined by the attributes that differentiate one from another, leading people to think more about differences between groups than similarities between groups. Obviously men and women differ in many ways, but the biggest differences are biological differences related to sex rather than psychological differences related to gender (Hyde, 2005). Stereotypes about gender that posit "essential" psychological differences between men and women, and even researchers who appeal to them (e.g., Baron-Cohen, 2003; Brizendine, 2007, 2011), appear to exaggerate the relatively small differences that exist. Indeed, our experiments show that actual gender differences across a variety of psychological measures were relatively meager. Exaggeration is therefore not as reliable for attributes that people do not perceive as "essential" for defining the difference between men and women (Experiment 2), or for known individuals within a group (Experiment 3). The magnitude of stereotype exaggeration therefore relies heavily on expectations about group differences.

Our experiments also provide weak hints to other processes that may contribute to stereotype exaggeration. In four out of the eight comparisons in which we tested participants' predictions about men's and women's performance, women exaggerated performance differences more than men. This finding may suggest that exaggeration was partly driven by ingroup favoritism (or valence inaccuracy; Judd & Park, 1993), sometimes predicting that one's own group would perform better on a desirable trait than another group. Interestingly, we did not observe similar overestimation by women of their personal performance (see Tables 1-3), consistent with predictions of individuated targets. In fact, in none of the seven comparisons in which we examined gender differences in self-performance predictions did women significantly predict they performed better than men. Overall,

these findings may imply that exaggeration may be weakly, and inconsistently, attenuated when perceivers are not members of a group that has an advantage in terms of the evaluated trait (e.g., men when the evaluated trait is social sensitivity ability). However, because overall, the gender effect in prediction was weak ($r = .16$) and inconsistent across experiments, these suggestions await further research.

Although our experiments test the hypotheses about stereotype exaggeration by studying gender stereotypes, we believe the mechanism underlying our account should account for the magnitude of stereotype exaggeration in any domain, and can even apply to people's beliefs about themselves. Groups, as well as oneself, are defined by the features that make a group or a person distinct from others. The most distinguishing features that define the core content of a group, or a self, those that differ from others, are therefore most likely to exaggerate differences between groups or between self and others. Groups seem more different from each other than they actually are (i.e., stereotype exaggeration). When applied to the self, it can make a person seem more different from others than he or she actually is (i.e., false uniqueness; Goethals, Messick, & Allison, 1991). People's beliefs about themselves and others are neither completely mistaken nor perfectly calibrated. Considering the process by which groups, or the self, are defined can provide a more precise understanding of the facts and fictions that exist in people's beliefs about a complicated world.

Appendix

Judd and Parks's (1993) conceptualization of stereotypic inaccuracy serves as the standard analysis approach in the field today, but their focus is different than ours because it focuses on estimates of attribute prevalence in single groups compared with a criterion measure (e.g., men, women) rather than on differences between groups (e.g., men vs. women). Their analysis approach therefore cannot test our hypotheses. However, we include it here for interested readers, as it does identify a potentially intriguing, unpredicted, and relatively consistent result. In particular, men's performance was consistently underestimated compared with their actual score on tests of social intellect, whereas predictions of women's performance were relatively calibrated. We present those analyses here, and discuss a mechanism that could explain it in the "General Discussion" section.

To perform this analysis, we computed the difference between predicted and actual performance for men and for women separately. These difference scores were then submitted to an ANOVA with participant's gender (men vs. women) as a between subjects variable and target's gender (men vs. women) as a within subjects variable.

In Experiments 1a through 1c, the grand means were all significantly larger than zero (1.67, 3.07, and 3.68, respectively), all $p < .001$, indicating that participants generally answered more items correctly than predicted. That is, participants

underestimated others' performance on these tests. The ANOVA also yielded a significant main effect for target's gender in all three experiments: Experiment 1a, $F(1, 231) = 109.70$, $p < .001$, $r = .57$; Experiment 1b, $F(1, 229) = 28.62$, $p < .001$, $r = .33$; and Experiment 1c, $F(1, 235) = 209.49$, $p < .001$, $r = .69$. This indicates that men's performance was underestimated more than women's performance. This analysis also yielded a significant target gender by participant gender interaction in Experiment 1a, $F(1, 231) = 5.20$, $p = .023$, $r = .15$; in Experiment 1b, $F(1, 229) = 6.76$, $p = .010$, $r = .17$; but not in Experiment 1c, $F(1, 235) = 2.03$, $p = .156$, $r = .09$. These interactions indicate that women predicted larger gender differences than men did on the DANVA2 and The ME test, but not on the EQ.

In Experiment 2, the grand means for the DANVA2, self-esteem, and happiness all differed significantly from zero (2.32, -10.58, and 2.16, respectively, all $ps < .001$). Participants thought others scored lower on the DANVA2 and happiness than others actually did, but they thought others scored higher on self-esteem than others actually did. In addition, the ANOVAs yielded a significant main effect of target's gender for the DANVA2, $F(1, 130) = 15.93$, $p < .001$, $r = .33$; and for self-esteem, $F(1, 130) = 63.44$, $p < .001$, $r = .57$; but not for happiness, $F(1, 130) < 1$. This indicates greater underestimation of men's performance than of women's performance on social sensitivity, and a great overestimation of men's scores on self-esteem scale than of women's scores. Finally, this analysis also yielded a significant target gender by participant gender interaction for the DANVA2, $F(1, 130) = 11.96$, $p = .001$, $r = .29$; a marginally significant interaction for self-esteem, $F(1, 130) = 3.71$, $p = .056$, $r = .17$; and no interaction for happiness, $F(130) < 1$. These interactions indicate that women predicted larger gender differences than men did on the DANVA2 and self-esteem, but not on happiness.

In Experiment 3, the grand mean was again significantly greater than zero (2.2), $t(179) = 6.53$, $p < .001$, indicating that participants underestimated others' performance on the DANVA2. This analysis also yielded a significant main effect for target gender, $F(1, 88) = 41.48$, $p < .001$, $r = .57$, indicating greater underestimation of men's performance (-3.65) than women's performance (-0.77). The analysis did not yield a significant target gender by participant gender interaction, $F(1, 88) = 2.48$, $p = .119$, $r = .16$, indicating that women did not predict larger gender differences than men did.

Given that our theory makes no predictions about these effects, and that overestimating or underestimating performance compared with an objective criterion is difficult to interpret, we do not speculate about these effects further.

Acknowledgments

The authors thank Jesus Diaz, Jasmine Kwong, Rachel Meng, Sarah Molouki, Michael Pang, Alyssa Pappas, Maimouna Thioune, Whitney Westmorland, and Emily Wolodiger for assistance conducting these experiments. They also thank Haotian Zhou, Michael Gilead, and Yoav Bar-Anan for their statistical advice.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Participants predicted their own performance in all experiments before predicting the performance of men and women. Because this is not the focus of our research and we did not have a specific prediction regarding these self-predictions, we report the data only in the tables and not in the text.
2. In Experiments 1 and 2, the order of predictions of men's and women's performance was counterbalanced.
3. Because the correlation between low self-esteem and high self-esteem (reverse coded) was low, we also analyzed the two self-esteem measures separately. We conducted a regression analysis predicting exaggeration on self-esteem by high self-esteem accessibility (reverse coded) and low self-esteem accessibility score controlling for social sensitivity accessibility. The regression yielded a significant effect for high self-esteem (reverse coded), $\beta = -.24$, $t(128) = -2.82$, $p = .006$, and for low self-esteem, $\beta = -.24$, $t(128) = -2.79$, $p = .006$, but not for social sensitivity, $\beta = -.13$, $t(128) = -1.51$, $p = .133$.
4. In Experiment 3, we did not run the additional bootstrapping procedure we report in Experiments 1 and 2, using only the first prediction participants gave, because in this experiment, we did not counterbalance the order of predicted performance of men and women.

Supplemental Material

Supplementary material is available online with this article.

References

- Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison Wesley.
- Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1-35). Hillsdale, NJ: Lawrence Erlbaum.
- Ashton, M. C., & Esses, V. M. (1999). Stereotype accuracy: Estimating the academic performance of ethnic groups. *Personality and Social Psychology Bulletin*, 25, 225-236.
- Baron-Cohen, S. (2003). *The essential difference: Men, women and the extreme male brain*. London, England: Penguin.
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorder*, 34, 163-175.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the eyes" test revised version: A study with normal adults, and adults with Asperger Syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42, 241-252.

- Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles, 41*, 279-296.
- Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles, 32*, 79-90.
- Brizendine, L. (2007). *The female brain*. New York, NY: Harmony.
- Brizendine, L. (2011). *The male brain*. New York, NY: Random House.
- Chambers, J. R., & Melnyk, D. (2006). Why do I hate thee? Conflict misperceptions and intergroup mistrust. *Personality and Social Psychology Bulletin, 32*, 1295-1311.
- Corneille, O., & Judd, C. M. (1999). Accentuation and sensitization effects in the categorization of multi-faceted stimuli. *Journal of Personality and Social Psychology, 77*, 927-941.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*, 177-193.
- Dawes, R. M., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude. *Journal of Personality and Social Psychology, 21*, 281-295.
- Diekmann, A. B., Eagly, A. H., & Kulesa, P. (2002). Accuracy and bias in stereotypes about the social and political attitudes of women and men. *Journal of Experimental Social Psychology, 38*, 268-282.
- Edwards, A. W. F. (1993). Galton, Karl Pearson and modern statistical theory. In M. Keynes (Ed.), *Sir Francis Galton, FRS: The legacy of his ideas* (pp. 91-107). Basingstoke, UK: Macmillan.
- Eiser, J. R. (1971). Enhancement of contrast in the absolute judgment of attitude statements. *Journal of Personality and Social Psychology, 17*, 1-10.
- Eyal, T., & Epley, N. (2010). How to seem telepathic: Enabling mind reading by matching construal. *Psychological Science, 21*, 700-705.
- Fiske, S. T., & Taylor, S. E. (1984). *Social cognition*. New York, NY: Random House.
- Ford, T. E., & Stangor, C. (1992). The role of diagnosticity in stereotype formation: Perceiving group means and variances. *Journal of Personality and Social Psychology, 63*, 356-367.
- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic accuracy model. *Psychological Review, 102*, 652-670.
- Gilbert, D. T. (1998). Ordinary personality. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 89-150). New York, NY: McGraw-Hill.
- Goethals, G. R., Messick, D. M., & Allison, S. T. (1991). The uniqueness bias: Studies of constructive social comparison. In J. Suls & T. A. Wills (Eds.), *Social comparison: Contemporary theory and research* (pp. 149-176). Hillsdale, NJ: Lawrence Erlbaum.
- Graham, J., Nosek, B. A., & Haidt, J. (2012). The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. *PLoS ONE, 7*, e50092.
- Gray, J. (1992). *Men are from Mars, women are from Venus: A practical guide for improving communication and getting what you want in your relationships*. New York, NY: HarperCollins.
- Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin, 131*, 898-924.
- Harrison, J. R., & Bazerman, M. H. (1995). Regression to the mean, expectation inflation and the winner's curse in organizational contexts. In R. M. Kramer & D. M. Messick (Eds.), *Negotiation as a social process: New trends in theory and research* (pp. 69-94). Thousand Oaks, CA: SAGE.
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581-592.
- Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review, 100*, 109-128.
- Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology, 61*, 366-379.
- Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. New York, NY: Oxford University Press.
- Jussim, L., Crawford, J. T., & Rubenstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science, 24*, 490-497.
- Krueger, J. I., Hasman, J. F., Acevedo, M., & Villano, P. (2003). Perceptions of trait typicality in gender stereotypes: Examining the role of attribution and categorization processes. *Personality and Social Psychology Bulletin, 29*, 108-116.
- Krueger, J. I., Savitsky, K., & Gilovich, T. (1999). Superstition and the regression effect. *Skeptical Inquirer, 23*, 24-29.
- Lyubomirsky, S., & Lepper, H. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46*, 137-155.
- Martin, C. L. (1987). A ratio measure of sex stereotyping. *Journal of Personality and Social Psychology, 52*, 489-499.
- McCauley, C., Thangavelu, K., & Rozin, P. (1988). Sex stereotyping of occupations in relation to television representations and census facts. *Basic and Applied Social Psychology, 9*, 197-212.
- McCrae, R. R., Chan, W., Jussim, L., DeFruyt, F., Lockenhoff, C. E., De Bolle, M., . . . Terracciano, A. (2013). The inaccuracy of national character stereotypes. *Journal of Research in Personality, 47*, 831-842.
- McGuire, W. J., & McGuire, C. V. (1988). Content and process in the experience of self. *Advances in Experimental Social Psychology, 21*, 97-144.
- Mussweiler, T. (2002). "Everything is relative": Comparison processes in social judgment. *European Journal of Social Psychology, 33*, 719-733.
- Nelson, L. J., & Miller, D. T. (1995). The distinctiveness effect in social categorization: You are what makes you unusual. *Psychological Science, 6*, 246-249.
- Nowicki, S., & Duke, M. P., Jr. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior, 18*, 9-36.
- Quellar, S., Schell, T., & Mason, W. (2006). A novel view of between categories contrast and within-categories assimilation. *Journal of Personality and Social Psychology, 9*, 406-422.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Stangor, C., & Schaller, M. (1996). Stereotypes as individual and collective representations. In C. N. Macrae, M. Hewstone, & C. Stangor (Eds.), *Foundations of stereotypes and stereotyping* (pp. 3-37). New York, NY: Guilford Press.
- Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology, 54*, 101-114.
- Vazire, S. (2010). Who knows what about a person? The Self-Other Knowledge Asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 28*, 281-300.

Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., . . . Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences, 41*, 929-940.

Wolsko, C., Park, B., Judd, C. M., & Wittenbrink, B. (2000). Framing interethnic ideology: Effects of multicultural and color-blind perspectives on judgments of groups and individuals. *Journal of Personality and Social Psychology, 78*, 635-654.