Statements of equivalence can imply differences: Asymmetries in directional comparisons

Eleanor K. Chestnut (ekc@stanford.edu)
Carla Remulla (remulla@stanford.edu)
Ellen M. Markman (markman@stanford.edu)
Department of Psychology, Stanford University, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Directional comparisons are often used to express similarity (e.g., "North Korea is like China"). These statements, however, frame the subject as the less typical *figure* and the complement as the more typical or prominent *ground*. Thus, despite expressing similarity, directional comparisons may imply that the *ground* is more representative. In Study 1, we analyze Twitter to show that directional comparisons occur in everyday conversation about gender; that men are the *ground* more often than women; and that only males frequently serve as the *ground* for positive traits (e.g., "Girls are as smart as boys"), suggesting that positive traits are considered typical of males, but not females. In Study 2, we show that directional comparisons intended to express equivalent ability (e.g., "Boys are as good as girls at a game called gorp") cause adults to infer that the *ground* has more natural skill and that the *figure* has to work harder.

Keywords: language, comparison, pragmatics, gender

Introduction

A group of middle school girls participating in a computer science program was recently told, "You're as smart as any boy in the world." Despite the clear intention of this statement to express gender equivalence, research dating back to Tversky's (1977) seminal work on the linguistic framing of similarity suggests that comparing one group *to* another may not actually convey true equality (e.g., Gleitman, Gleitman, Miller, & Ostrin, 1996; Bruckmüller & Abele, 2010; Chestnut & Markman, 2014). Rather, directional comparisons frame the item in the subject position as the less typical *figure* and the item in the complement position as the more typical or prominent *ground*. In Tversky's (1977) original study, for example, adults preferred to state, "North Korea is similar to China," rather than, "China is similar to North Korea," because at the time, China was a more prominent and well-known country, and therefore served as a better *ground* than North Korea. Thus, items framed as *grounds* in comparisons may be inferred to be more typical than items framed as *figures*, even if the sentence expresses similarity.

Since typicality is often conflated with positive traits such as high status and prestige (Eagly & Kite, 1987; Devos & Banaji, 2005; Hegarty & Bruckmüller, 2013), groups or individuals framed as *grounds* may also be associated with higher status. Gleitman et al. (1996) provided evidence for this association by showing that when presented with sentences containing novel words such as, "The zum is similar to the gax," adults inferred that the *gax* was more important and famous than the *zum*. Since adults had no prior experience with these novel words, their inference was based purely on the direction of the comparison. Similarly, Chestnut and Markman (2014) demonstrated that after hearing sentences such as, "The blicket plays soccer as well as the toma," children ages six to eight inferred that the *toma* was the older and more established player. Stating, "Girls are as good as boys at math," then, may subtly imply that boys are the typical, higher-status, or even more naturally skilled mathematicians.

To date, no research has empirically tested this inferential process for statements expressing similarity between familiar groups (e.g., "Women are as smart as men"), or how such statements occur in everyday discourse. In the following studies, we use comparisons between genders as a case study for answering these questions.

To investigate adults' use of directional comparisons for expressing gender similarities, in Study 1 we analyzed public messages ("tweets") on the microblog website Twitter, which has recently been recognized as a reliable way of measuring linguistic trends (Cook, Han, & Baldwin, 2014; Doyle, 2014; Eisenstein, 2014). Twitter is a particularly interesting corpus to use for two reasons: 1) it is a free website with over 280 million users who post approximately 500 million tweets per day, making it a large source of natural language data; and 2) it allows users to "follow" others (i.e., view others' tweets on their home pages), which suggests that users typically compose tweets with the expectation that others will read them ("About Twitter", 2015). Thus, Twitter provides an opportunity to observe how adults spontaneously frame gender similarities in naturalistic, contemporary conversation.

In Study 2, we provide experimental evidence about whether such comparisons can shape adults' beliefs about the groups being compared. Specifically, we ask whether directional comparisons intended to express equivalent ability between genders at a novel activity (e.g., "Girls are as good as boys at a game called gorp") nevertheless cause adults to associate greater natural skill with the gender framed as the *ground* and greater effort with the gender framed as the *figure*.

Study 1

In Study 1, we used a corpus of public tweets to measure the relative frequencies of comparisons expressing gender

similarities in everyday language. Specifically, we asked whether the phrases "women are as * as men" and "girls are as * as boys", in which the asterisk could be replaced with any word, occurred more or less often than "men are as * as women" and "boys are as * as girls", respectively, over the past year.

We also asked whether the valence of the words that replace the asterisks change when females rather than males serve as the *ground* (e.g., "Men are as *catty* as women"; "Women are as *capable* as men"). If men serve as the *ground* for positive traits more often than women, for example, then this would suggest that people view positive traits as more typical of men.

Methods

Twitter Extraction One year (11/29/2013-11/29/2014) of public Twitter tweets were sampled. Data were extracted by typing a string-matching pattern into the search bar of Twitter.com, and scrolling down to retrieve all search results from the year of interest. The search patterns used were "girls are as * as boys", "boys are as * as girls", "women are as * as men", and "men are as * as women". Each search returned public tweets posted by users, where the asterisk matched a single word. The total number of tweets returned was 2370. All searches were conducted on the same day, 11/30/2014.

The html from the browser in which each search was performed was saved. A python html parser utilizing the package BeautifulSoup was used to extract the tweet text and the target word matching the asterisk for each tweet. A complete list of all tweets can be found at http://web.stanford.edu/~gordonam/Publications_files/tweetsSample1908.csv.

Equivalence Rating In some cases, the search patterns returned tweets that did not actually express gender equivalence (e.g., "I don't think that women are as smart as men"). To obtain a sample of tweets in which all tweets expressed equivalence between genders, three coders separately coded each tweet for whether it expressed equivalence (1) or non-equivalence (0). For this coding process, all gendered terms (e.g., *women*) were removed from the tweets and replaced with "A" (for *figures*) and "B" (for *grounds*; e.g., "A are as smart as B"). Tweets were coded as not expressing equivalence if the phrase was negated (e.g., "I don't think that A are as smart as B"), if the phrase was hypothetical (e.g., "If A were as smart as B, the world would be a better place"), or if the phrase did not actually have the syntax of interest (e.g., "A are as harsh as B are vindictive"). There was strong agreement among the three coders (Cronbach's α = .91). For a tweet to be included in the final sample, all three coders had to agree that the tweet expressed equivalence between A and B. This resulted in a final sample of 1908 tweets.

Valence Rating Valence ratings were performed by 10 raters blind to the hypothesis of the study. Each rater was shown all 243 unique target words (i.e., the words replacing the asterisks) from the sample of 1908 tweets. Raters viewed one word per trial, and word order was randomized for each rater. Raters were instructed to adjust a slider on a scale of -100 (negative) to 100 (positive) to indicate how negative or positive they perceived the word to be when describing a person. Cronbach's alpha for this task was .98, demonstrating that there was strong agreement across raters for each word.

In some cases, the valence of the target word in isolation differed from the valence of the word in the context of the sentence (e.g., "Women are as *responsible* as men for violence"). In other cases, the target word was not an adjective (e.g., "Boys are as *likely* as girls to be unhappy with their bodies"). To estimate the prevalence of these cases, two coders identified target words that had a contextual valence mismatch or were non-adjectives in a random sample of 200 tweets. There was 100% agreement between the coders, who identified 3 mismatches and 2 non-adjective target words (1.5% and 1% of the sample, respectively). Since these numbers were very low, we did not remove these tweets from our sample.

Results and Discussion

Our searches on Twitter yielded a total of 1908 gendered comparisons over the course of one year.

We first tested for differences between the number of comparisons framing boys or men as the *ground* ("girls are as * as boys"; "women are as * as men") and the number of comparisons framing girls or women as the *ground* ("boys are as * as girls"; "men are as * as women"), respectively. We found that the number of comparisons framing men as the *ground* (711) was significantly greater than the number of comparisons framing women as the *ground* (411; binomial sign test, $p < .001$). Since the *ground* position is associated with higher status (e.g., Bruckmüller & Abele, 2010), this result arguably reflects the perception of men as the higher-status gender (e.g., Eagly & Wood, 1982).

We did not find a difference between the number of comparisons framing boys (380) and girls (406) as the *ground* (binomial sign test, $p = .37$), suggesting that boys are not favored over girls as the *ground* of gendered comparisons the way that men are favored over women.

To determine whether comparisons framing females as the *ground* and comparisons framing males as the *ground* differed in valence, we compared the means of the distributions of valence for each gender, within each age group (women/men or girls/boys). Because these distributions were non-normal (see Figure 1), we assessed the confidence intervals using bootstrapping, and the p-values with permutation testing. We found differences in valence between comparisons framing women (M = -38.27, SD = 26.74) and men (M = -3.12, SD = 58.98) as the *ground* ($p < .001$), such that men served as the *ground* for more positively valenced traits. In terms of frequency, women served as the *ground* for 378 negative comparisons (values -100 to 0) and 33 positive comparisons (values 0 to 100),

while men served as the *ground* for 350 negative comparisons and 361 positive comparisons. A chi-squared test comparing the number of positive and negative comparisons when women and men served as the *ground* also showed that men served as the *ground* for positive traits more often than did women, $\chi^2(1) = 206.98$, $p < .001$.

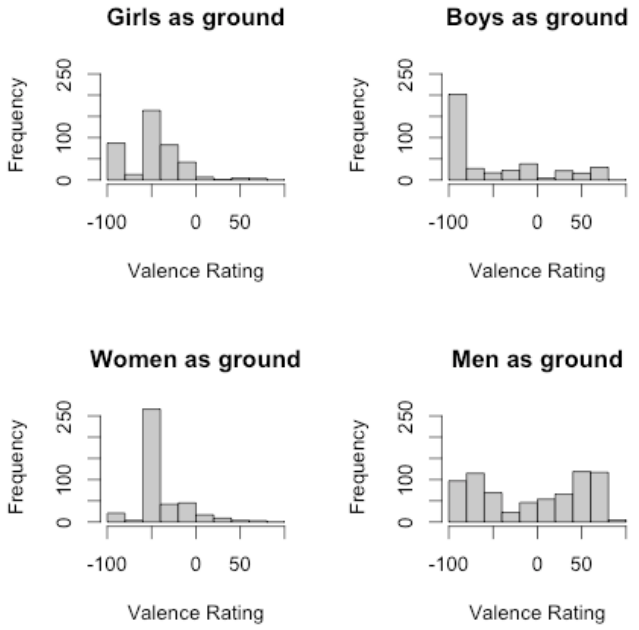

Figure 1: Histograms of valence ratings when girls, boys, women, and men served as the *ground*.

We did not find differences in valence between comparisons framing girls (M = -44.95, SD = 29.79) and boys (M = -44.11, SD = 53.34) as the *ground* (p = .81), but as Figure 1 shows, the distributions when girls and boys served as the *ground* were significantly different (D = .38, p < .001). A chi-squared test comparing the number of positive and negative comparisons when girls and boys served as the *ground* showed that, like men, boys served as the *ground* for positive comparisons more often ($\chi^2(1)$ = 42.22, p < .001). While girls served as the *ground* for 389 negative comparisons and 17 positive comparisons, boys served as the *ground* for 307 negative comparisons and 73 positive comparisons.

The difference between genders, then, is that boys and men tended to serve as the *ground* for both positive and negative traits while girls and women served as the *ground* primarily for negative traits.

Study 2

In our second study, we explored a possible implication of this asymmetry for reasoning about ability. We focused on one particular way of expressing equivalent skill, namely: "A is *as good as* B". Consider, for example, the statement: "Girls are as good as boys at math." Such statements, by implying that boys are more typically good at math, may further suggest that boys are more naturally skilled at math while girls have to work harder to succeed.

We were interested in the concepts of natural skill and effort for two reasons. First, natural skill and effort are sometimes seen as diametrically opposed. If a person has to work hard to succeed at an activity, for example, then that person might be viewed as lacking natural talent (e.g., Tiedemann, 2000; Dweck, 2007). Groups frequently framed as *figures* in comparisons, then, may be associated strongly with effort if groups framed as *grounds* are associated with natural skill. Second, high status, a feature of the *ground* position, can be conflated with natural skill (e.g., Leslie, Cimpian, Meyer, & Freeland, 2015).

We presented participants with sentences containing the structure, "A is as good as B at X," where A and B were two genders (girls and boys or women and men), and X was a novel activity (e.g., a dance called *quibbing*). We then asked participants one of two questions: 1) Who do you think has to work harder to be good at X? (*Effort* condition, Study 2A), or 2) Who do you think is more naturally skilled at X? (*Skill* condition, Study 2B). We compared these conditions to *Baseline* conditions, in which we measured participants' preexisting gender biases. We predicted that the gender serving as the *figure* (the subject) would be assumed to have to work harder than it would otherwise (Study 2A), while the gender serving as the *ground* (the complement) would be inferred to have greater natural skill than it would otherwise (Study 2B).

Study 2A

We first investigated whether directional statements intended to express equality between genders in a given activity (e.g., "Men are as good as women at a game called gorp") increase the assumption, relative to baseline, that the gender framed as the *figure* (men, in this case) has to work hard to have that level of ability.

Methods

Participants Participants were 240 English-speaking adults in the US ages 18 to 66 (M = 32, 116 men) who participated on Amazon Mechanical Turk for a payment of $0.15. A total of 160 participants participated in the *Effort* condition (M = 32, 18-60, 72 men), and 80 participants participated in the *Effort Baseline* condition (M = 33, 19-66, 44 men).

Materials Four sentences were used in each condition. In the *Effort* condition, each sentence compared either girls' and boys' or women's and men's ability to do one of four novel activities, taken from Cimpian and Markman (2011): a dance called *quibbing*, a puzzle called *zool*, a sport called *leeming*, and a game called *gorp*. The structure of all four sentences was the same: "Gender A is as good as Gender B at X" (e.g., "Boys are as good as girls at a dance called quibbing"). In the *Effort Baseline* condition, the four sentences simply stated that the activity existed (e.g., "There exists a dance called quibbing").

Procedure Participants participated in one of two conditions: the *Effort* condition or the *Effort Baseline* condition. In the *Effort Baseline* condition, participants were presented with four sentences (e.g., "There exists a dance called quibbing"), one at a time. After each sentence was presented, participants were asked, "Who do you think has to work harder to be good at [quibbing]?" Participants indicated their responses by selecting one of the two genders. While half of the participants was asked to decide between girls and boys across all four trials, the other half was asked to decide between women and men. The order of the four trials was counterbalanced across participants, and the order of the two genders that participants were asked to choose from was counterbalanced both across and within participants. After participating, participants received a debriefing message explaining that in real life, both genders are good at the same activities.

The *Effort* condition was identical to the *Effort Baseline* condition, except that the sentences used were different, as described above, and required further counterbalancing. Each participant in the *Effort* condition was presented with two sentences that framed either girls or women as the *figure* and two sentences that framed either boys or men as the *figure*. For half of the participants, girls or women were framed as the *figure* for the game and the puzzle, and for the other half of the participants, girls or women were framed as the *figure* for the sport and the dance. To ensure that an equal number of participants viewed each sentence, we therefore included twice as many participants in the *Effort* condition as in the *Effort Baseline* condition.

Statistical Analysis To assess the effect of variables of interest on responses in Studies 2A and 2B, we ran mixed effects logistic regression models using the lme4 package of R (Bates, Maechler, Bolker, & Walker, 2014). To test for the significance of effects, we performed likelihood ratio tests. Chi-squared values, degrees of freedom, and *p*-values, all from the likelihood ratio test, are reported for each statistical test.

Results and Discussion

The dependent measure was the proportion of times participants stated that men or boys needed to expend more effort. We predicted that this number would be greater when men or boys served as the *figure* of the comparison (e.g., "Men are as good as women at a game called gorp") than when they served as the *ground* (e.g., "Women are as good as men at a game called gorp"), or when there was no comparison (e.g., "There exists a game called gorp").

We fit the data using a mixed effects logistic regression model with fixed effects of participant gender (female or male) and *figure* (female, male, or none) and a random effect of item. Including age group (boys/girls or men/women) as a fixed effect did not improve model fit ($\chi^2(1) = .04$, $p = .84$), nor did including an interaction between age group and *figure* ($\chi^2(3) = 6.04$, $p = .11$), so we did not model the effect of age group in further analyses.

The effect of participant gender was marginal, such that, overall, female participants ($M = .55$, $SE = .04$, $n = 124$) were more likely than male participants ($M = .49$, $SE = .05$, $n = 116$) to state that males had to work harder to be good at the activity ($\chi^2(1) = 2.80$, $p = .09$). This suggests that a person's gender may influence how that person perceives gender categories, biasing them (at least in the case of women, whose mean differed more from chance) towards thinking more positively of their own gender.

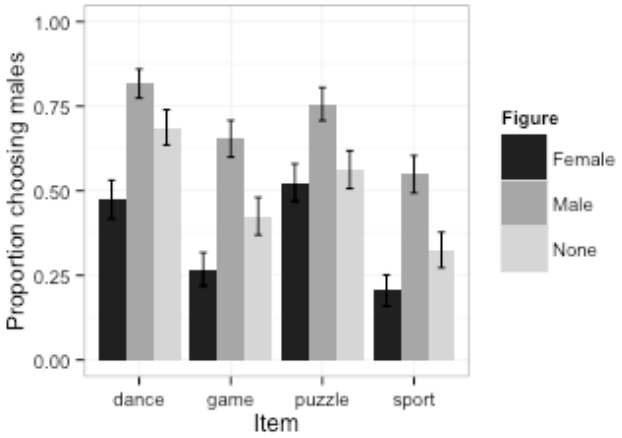

Figure 2: The proportion of participants choosing males when the *figure* was female or male (*Effort* condition), or when there was no *figure* (*Effort Baseline* condition). Error bars represent +/- 1 SEM.

Participants responded to the question, "Who do you think has to work harder to be good at [quibbing]?" by more frequently selecting the *figure*, or subject, of the comparison ($\chi^2(2) = 74.5$, $p < .001$). In follow-up analyses to determine whether this effect held when both females and males served as the *figure*, we tested planned contrasts between the *Effort* and *Effort Baseline* conditions. These analyses showed that, as predicted, when the *figure* was female (e.g., "Girls are as good as boys at a dance called quibbing"), participants were less likely to state that males had to work harder to be good ($M = .37$, $SE = .03$) than at baseline ($M = .50$, $SE = .03$; $\chi^2(1) = 12.86$, $p < .001$). Similarly, when the *figure* was male (e.g., "Boys are as good as girls at a dance called quibbing"), participants were *more* likely to state that males had to work harder to be good ($M = .69$, $SE = .03$) than at baseline ($\chi^2(1) = 24.56$, $p < .001$; see Figure 2).

Directional comparisons, then, shape the way that adults interpret a gender's ability to do an activity. On the surface, stating that girls are as good as boys at a game seems to convey equality between girls and boys, but it actually implies that girls, and not boys, have to work hard to have that level of ability.

Study 2B

In this study we investigated whether adults generate inferences based on the *ground* position, as well. Using the

same sentences used in Study 2A (e.g., "Women are as good as men at a dance called quibbing"), we asked whether adults associate more raw, natural skill with the gender framed as the *ground* of a comparison (here, men) than they would typically associate with that gender.

Methods

Participants Participants were 240 English-speaking adults in the US ages 18 to 74 (M = 30, 147 men) who participated on Amazon Mechanical Turk for a payment of $0.15. A total of 160 participants participated in the *Skill* condition (M = 29, 18-74, 102 men), and 80 participants participated in the *Skill Baseline* condition (M = 32, 19-70, 45 men).

Materials The materials were identical to those used in Study 2A.

Procedure The procedure was identical to study 2A, except that questions targeted natural skill rather than effort (e.g., "Who do you think is more naturally skilled at quibbing?"). Thus, participants participated in either the *Skill* condition or the *Skill Baseline* condition.

Results and Discussion

To assess whether directional comparisons intended to express equivalent ability further imply that the group in the *ground* position has more natural skill, we now asked, "Who do you think is more naturally skilled at [gorp]?"

The dependent measure was the proportion of times participants stated that men or boys had more natural skill. We predicted that this number would be greater when men or boys served as the *ground* of the comparison (e.g., "Girls are as good as boys at a game called gorp") than when they served as the *figure* (e.g., "Boys are as good as girls at a game called gorp"), or when there was no comparison (e.g., "There exists a game called gorp"). We fit the data using a mixed effects logistic regression model with fixed effects of participant gender (female or male) and *ground* (female, male, or none) and a random effect of item. Including age group (boys/girls or men/women) as a fixed effect did not improve model fit ($\chi^2(1) = .49$, $p = .48$), nor did including an interaction between age group and *ground* ($\chi^2(1) = 4.66$, $p = .20$), so we did not model the effect of age group in further analyses.

The effect of participant gender was significant ($\chi^2(2) = 9.37$, $p = .002$). Overall, female participants ($M = .41$, $SE = .05$, $n = 93$) were less likely than male participants ($M = .50$, $SE = .04$, $n = 147$) to state that males were more naturally skilled at the activity. This result, together with the marginal effect of participant gender in Study 2A, suggests that women may be biased towards answering questions such as these with their own gender, or towards attributing greater natural ability to females rather than males.

Participants responded to the question, "Who do you think is more naturally skilled at [quibbing]?" by more frequently selecting the *ground*, or complement, of the comparison ($\chi^2(2) = 169.34$, $p < .001$). In follow-up analyses to determine whether this effect held when both females and males served as the *ground*, we tested planned contrasts between the *Skill* condition and the *Skill Baseline* condition. These analyses showed that, as predicted, when the *ground* was female (e.g., "Boys are as good as girls at a dance called quibbing"), participants were less likely to state that males were more naturally skilled ($M = .21$, $SE = .02$) than at baseline ($M = .48$, $SE = .03$; $\chi^2(1) = 49.78$, $p < .001$). When the *ground* was male (e.g., "Girls are as good as boys at a dance called quibbing"), participants were *more* likely to state that males were more naturally skilled ($M = .70$, $SE = .03$) than at baseline ($\chi^2(1) = 32.47$, $p < .001$; see Figure 3).

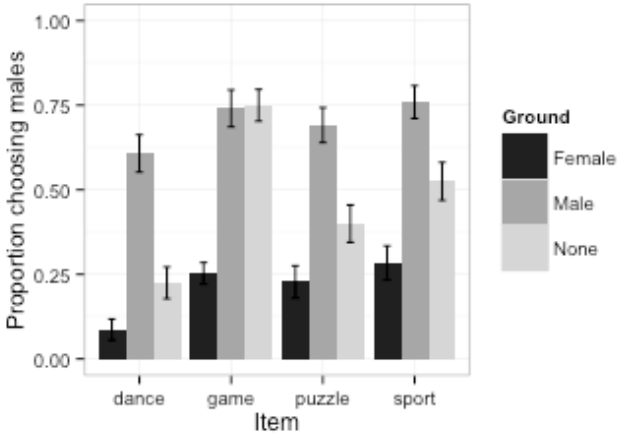

Figure 3: The proportion of participants choosing males when the *ground* was female or male (*Skill* condition), or when there was no *ground* (*Skill Baseline* condition). Error bars represent +/- 1 SEM.

Just as the *figure* position was associated with having to put forth more effort to succeed at an activity (Study 2A), so the *ground* position was associated with greater raw talent.

General Discussion

This work explored the use of directional comparisons (e.g., "Women are as smart as men") in everyday communication (Study 1) and the subtle differences in ability that such comparisons imply (Study 2).

In Study 1, we showed, using gender as a case study, that directional comparisons are common ways of expressing similarity between groups, and that the way adults frame these comparisons is non-arbitrary. While males, especially men, frequently served as the *ground* for positive traits (e.g., "Women are as smart as men"), females overwhelmingly served as the *ground* for negative traits (e.g., "Boys are as fake as girls"). We also found that men served as the *ground* significantly more often than did women.

Our Twitter results arguably reflect current stereotypic beliefs about gender. As Miller et al. (1991) showed, adults explain group differences by focusing on the less typical, "deviant" group, placing them in the subject, or *figure*,

position. When explaining gender gaps in voting, for example, adults in their study focused on how *women* differed from *men*, rather than the reverse, because typical voters were seen as male. Thus, Twitter users likely framed their comparisons according to whether they viewed the trait as typical of females or males, placing the gender most strongly associated with the trait in the *ground* position.

In addition to *reflecting* stereotypic beliefs about gender, we argue that these directional comparisons may also *contribute* to such beliefs. With respect to frequency, by framing men as the *ground* of comparisons more often than women, we may be *strengthening* the idea that men are the typical, higher-status gender, even when the comparison does not explicitly concern status or power. Bruckmüller, Hegarty, & Abele (2012) provide some evidence for this by showing that adults who read paragraphs describing how women's leadership styles differ from men's were more likely than those who read paragraphs describing how men's leadership styles differ from women's to endorse the idea that men are more powerful and agentic.

Our own results in Studies 2A and 2B more clearly show that directional comparisons can exacerbate, in addition to reflecting, group biases. They also provide evidence for an inferential process that has not been investigated previously. In these studies, we found that, after reading comparisons intended to express equivalent ability between genders at a novel activity (e.g., "Boys are as good as girls at a game called gorp"), adults were more likely to state that the gender framed as the *figure* (here, boys) had to work hard to be good at the activity and that the gender framed as the *ground* (here, girls) was more naturally skilled. Thus, directional comparisons may subtly suggest that the group framed as the *ground* is the group with raw talent, and that equivalent ability between the two groups has been achieved only because the group framed as the *figure* has put forth enough effort.

Stating that girls are as good as boys at math, then, may implicitly perpetuate gender stereotypes by strengthening both the idea that girls have to work hard to be good at math and the idea that boys are naturally talented mathematicians.

Acknowledgments

Special thanks to Co A. Tran, Sarina M. Soriano-Taylor, and Alan M. Gordon for their help coding and analyzing data from Twitter.

References

About Twitter, Inc. (2015). Retrieved February 1, 2015, from https://about.twitter.com/company

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed models using Eigen and S4. *R package 1.1-7, http://CRAN.R-project.org/package=lme4.*

Bruckmüller, S., & Abele, A. (2010). Comparison focus in intergroup comparisons: Who we compare to whom influences who we see as powerful and agentic. *Personality and Social Psychology Bulletin, 36*(10), 1424-1435.

Bruckmüller, S., Hegarty, P., & Abele, A. (2012). Framing gender differences: Linguistic normativity affects perceptions of power and gender stereotypes. European Journal of Social Psychology, *42*(2), 210-218.

Chestnut, E., & Markman, M. (2014). Children's inferences based on figure and ground thematic roles. In P. Bello, M. Guarini, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 325-330). Austin, TX: Cognitive Science Society.

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development, 82*(2), 471-492.

Cook, P., Han, B., & Baldwin, T. (2014). Statistical methods for identifying local dialectal terms from GPS-tagged documents. *Dictionary: Journal of the Dictionary Society of North America*, *1*(35), 248-271.

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447-466.

Doyle, G. (2014). Mapping dialectal variation by querying social media. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 98-106). Gothenburg, Sweden: Association for Computational Linguistics.

Dweck, C. S. (2007). The secret to raising smart kids. *Scientific American Mind*, *18*, 36-43.

Eagly, A. H., & Wood, W. (1982). Inferred sex differences in status as a determinant of gender stereotypes about social influence. *Journal of Personality and Social Psychology*, *43*, 915-928.

Eagly, A. H., & Kite, M. E. (1987). Are stereotypes of nationalities applied to both women and men? *Journal of Personality and Social Psychology*, *53*, 451-462.

Eisenstein, J. (in press). Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *Handbook of Dialectology*. Wiley-Blackwell Press.

Gleitman, L., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts, *Cognition, 58*(3), 321-376.

Hegarty, P., & Bruckmüller, S. (2013). Asymmetric explanations of group differences: Experimental evidence for Foucault's disciplinary power. *Social and Personality Psychology Compass*, *7*(3), 176-186.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262-265.

Miller, D. T., Taylor, B., & Buck, M. L. (1991). Gender gaps: Who needs to be explained? *Journal of Personality and Social Psychology*, *61*(1), 5-12.

Tiedemann, J. (2000). Gender-related beliefs of teachers in elementary school mathematics. *Educational studies in mathematics*, *41*, 191-207.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.