PNAS

Supporting Information for

Gesture counteracts gender stereotypes conveyed through subtle linguistic cues

Yihan Qian[1], Susan Goldin-Meadow[1], Lin Bian[1]

[1]Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL, 60637

Corresponding authors: Yihan Qian, Lin Bian

Email: yihanq@uchicago.edu, linbian@uchicago.edu

This PDF file includes:

Supporting text

- Participants
- Extended methods
- Coding plan
- Analyses
- Supplemental analyses
- Final comments

SI References

Other Supplementary Materials for this manuscript include the following:

Study materials, data files and code can be found here: https://osf.io/gqypb/

Supporting Text

Participants

Study 1. In Study 1, we recruited 160 children (80 girls, 80 boys) aged 8 to 11 years (M = 9.98 years, SD = 1.16) to properly counterbalance all aspects of the study (i.e., gender, age, order of the questions, etc.). An *a priori* power analysis (G*Power 3.1; 1) for a regression model with condition (i.e., no gesture vs. *equal* gesture) as the only predictor was performed prior to the data collection. The power analysis suggested that a sample size of 152 participants would be required to detect a medium effect size (f^2 = 0.46), assuming a power of .80 and an alpha of .05. This age range was comparable to the range used in a previous study examining children's understanding of subtle linguistic phrasing (i.e., Subject Complement Statements, SCS); the study suggested that children by age 8 display recall and understanding of statements relevant to our research focus (2). An additional 27 participants were tested but not included in the final analyses due to failed attempts to pass the manipulation check (n = 20; see Extended Methods), unwillingness to complete the study (n = 3), and parental or sibling distractions (n = 2). Two other participants were not included in the final sample due to language difficulties (n = 2); this exclusion was not pre-registered but being able to fluently repeat the sentences was an important criterion for our study.

Study 2. Consistent with Study 1, we included 160 children (80 girls, 80 boys) aged 8 to 11 years (M = 9.98 years, SD = 1.16). An additional 21 participants were excluded from the sample because of failed attempts to pass the manipulation check (n = 17), technical difficulties (n = 2), and parental or sibling distractions (n = 2).

Across both studies, participants were recruited from the lab database (279 children; 87.2% of the participants) or a local museum (41 children; 12.8% of the participants) in the Midwestern United States. Ninety-two percent of families provided their children's race/ethnicity information. Of those, 47.8% identified as White, 14.7% as Asian, 13.8% as multiracial, 7.8% as Black, 7.5% as Latinx, 0.6% as Native Hawaiian/Pacific Islander, and 0.3% as Middle Eastern. Seventy-three percent of parents reported their family annual income; the median household income was $110,000. Ninety-one percent of parents reported parental education information, and within this subset, 75.0% of families had at least one parent with a bachelor's degree or higher.

Extended Methods

Study 1 (Pre-registration at: https://aspredicted.org/K76_F4H)

Each child was randomly assigned to one of the two conditions: Equal Gesture condition or No Gesture condition.

Practice trials. The procedures began with participants completing two practice trials (adapted from 2), designed to acquaint participants with the task and the scale. On each trial, a picture (e.g., a lemon and a banana) and a description sentence (e.g., "Here are a lemon and a banana.") were presented to the participants. The experimenter read the sentence to the child and asked the child to repeat it. Based on the child's answer, the experimenter provided feedback about the child's repetition (e.g., "That's right, / Actually, here are a lemon and a banana."). The child participant then answered a forced-choice question and a scale question. For example, the child indicated which fruit was sweeter and rated how much sweeter their selection was on a scale from *a little bit* (1) to *some* (2) to *a whole lot* (3). The order of the two practice trials was counterbalanced across participants.

Test trials. After the practice trials, participants were introduced to a fictional group from a faraway planet and completed four test trials in counterbalanced order. In each test trial, children watched a short video clip of an actress reading a subject-complement statement about girls and boys in a novel domain (e.g., "On this planet, girls are as good as boys at yuzzing").

Video stimuli. As described in the main text, children in the No Gesture condition watched the videos of the actress reading the SCS with her hands in the resting position, whereas children in the Equal Gesture condition watched the videos of the actress making the *equal* gesture while reading the SCS. Each test trial involved one of the four novel domains: hodding, zorbing, plipping, or yuzzing, to minimize the potential impact of pre-existing gender stereotypes on children's inferences. The order of gender groups in the statements was also counterbalanced across trials and participants (i.e., in total, participants heard girls placed in the subject position and boys placed in the complement position for two trials, and vice versa for the other two trials).

Manipulation check. A manipulation check was conducted after each video. Since our research question was whether the *equal* gesture can counteract the biasing message transmitted through subtle linguistic cues, being able to understand the syntactic structure of the SCS was a prerequisite for participating in the study. Sentence repetition is a sensitive measure of a child's language

skills because children often struggle to repeat sentences that they find difficult to understand (4, 5). To ensure that the children were able to understand the SCS, we followed the previous research (2) and asked children to repeat the SCS before comparing two gender groups. Specifically, after showing each video to a child, the experimenter gently asked the child to repeat what the adult said in the video clip. Children's repetitions were considered correct if the following two criteria were met. First, the order of the two gender groups in the repeated sentence matched the order in the video (i.e., which group was mentioned first). Second, the comparative phrase, "as good as" was produced in the repetition. As pre-registered, those who were not able to correctly repeat at least 2 out of 4 sentences were eliminated from the final analysis. Depending on whether the child successfully recalled the sentence, they were reinforced or corrected with the accurate statement (e.g., "That's right, / Actually, on this planet, girls are as good as boys at yuzzing.").

Stereotype measures. Next, participants answered two questions gauging their endorsement of gender stereotypes. First, we measured children's inference of gender stereotypes in a forced-choice format by asking either (1) which gender group was more **naturally skilled** (e.g., "Are girls naturally better at yuzzing, or are boys naturally better at yuzzing?"), or (2) which gender group had to **work harder** to obtain the skill (e.g., "Do girls need to work harder to be good at yuzzing, or do boys need to work harder to be good at yuzzing?). Participants were asked about natural ability in two trials and about effort in the other two trials. The mapping between question type and domain was counterbalanced across participants.

Second, the strength of participants' stereotypical inferences was assessed on a scale ranging from *a little bit* to *some* to *a whole lot* based on the gender group children selected in the first question (e.g., if a child chose boys as **naturally better** in the previous question, they were asked, "By how much do you think boys are naturally better than girls at yuzzing?" If a child chose girls as needing to **work harder** in the previous question, they were asked, "By how much harder do you think girls need to work than boys to be good at yuzzing?").

Debrief. After completing all four test trials, children received a debriefing explaining that the story was fictional and that everyone in real life can be good at any activity as long as they try really hard and practice a lot.

Study 2 (Pre-registration at https://aspredicted.org/ZSY_GTQ)

The overall procedure of Study 2 was largely identical to that of Study 1, with five exceptions. These changes (except for the last one) were intended to direct participants' attention to the experimenter's hands because the focus of Study 2 was to compare two gestural conditions (*unequal* vs. *equal* gesture).

First, in the practice task, participants were not presented with a picture stimulus or asked to repeat the descriptive sentence as in Study 1. Instead, participants watched a short video clip of an actress comparing two types of fruits while making the thumbs-up and thumbs-down emblems (i.e., "I am really into lemons <thumbs-up>, but I am not really into apples <thumbs-down>."). Participants were then asked whether they saw a thumbs-up or a thumbs-down when either lemons or apples were mentioned.

Second, unlike Study 1 where participants watched the video clip once, each video in Study 2 was played twice for the children.

Third, in addition to repeating the exact sentence after each manipulation check, the experimenter mimicked the actress from the video and made the same gesture while reading the SCS.

Fourth, as mentioned in the main text, to ensure consistency between the Unequal and Equal Gesture conditions in Study 2, rather than raising both hands simultaneously to shoulder level (as in Study 1), the *equal* gesture in Study 2 was produced by sequentially raising each hand to the same height. Although the two types of *equal* gestures in the two studies were produced differently with respect to simultaneity, the final display of both *equal* gestures was the same.

Lastly, we changed one of the novel activities in Study 1, zorbing, to giltching in Study 2, as we later discovered zorbing was not novel but a real-life activity that might be familiar to some participants.

Coding Plan

Forced-choice responses*.* Across the two studies, children received a score of 1 for a trial if they selected the reference group (i.e., the second-mentioned group, the group in the complement position) as the one with more natural ability, and 0 if they selected the variant group. Because children tend to reason about natural ability and effort as negatively related (3; also see Supplemental Analyses), when stereotype questions related to effort were asked, the coding was reversed. Overall, the participant received a score of 1 for each trial if the reference group was selected as naturally better or as requiring less effort, and a score of 0 if the variant group was chosen.

Weighted responses*.* The strength of participants' gender stereotypical inferences was rated on a six-point scale. The exact coding method for each trial depended on whether children received a question related to natural ability or effort, and the group children selected. In trials when stereotype questions related to natural ability were asked, children's scores ranged from 1 (the variant group was a whole lot better) to 6 (the reference group was a whole lot better).

The scoring system was reversed for stereotype questions related to effort. On each trial, children's scores ranged from 1 (the reference group needed to work a whole lot harder) to 6 (the variant group needed to work a whole lot harder). Overall, the higher the weighted score, the stronger the stereotypical belief participants held associating the reference group with more natural ability.

Analyses

We preregistered two primary analyses, identical for both Studies 1 and 2. First, we performed a mixed-effect logistic regression model on each participant's forced-choice response across each trial (coded as 1 if the reference group was selected as naturally better or required less effort in that trial; and 0 if the variant group was selected as naturally better or required less effort in that trial), including condition (Study 1: *equal* gesture present = 1, no gesture = -1; Study 2: *equal* gesture = 1, *unequal* gesture = -1), participant gender (female = 1, male = -1), age (continuous, in years), and all possible interactions among these factors as predictors, as well as a random intercept of participants. Although our preregistration specified the inclusion of "activity" as a random effect, the model fit was singular. Additionally, inspection of a model that only included "activity" as a random effect revealed that it contributed little variance—i.e., 0.000 in Study 1, and 0.004 in Study 2. Thus, "activity" was dropped as a random effect.

Second, we conducted a mixed effects linear regression model on weighted responses using the same fixed and random factors.

Study 1. In Study 1, a significant main effect of condition on participants' inferences was found for forced-choice responses. Compared to participants in the No Gesture condition, children in the Equal Gesture condition were less likely to attribute natural ability to the reference group, Wald $\chi2 = 11.65$, $p < .001$. Weighted scores also paralleled the pattern observed in forced-choice responses and together supported the *equal* gesture's contribution in mitigating gender stereotyping ($B = -0.38$, $SE = 0.16$, $p = .020$). We also found a main effect of age when forced-choice responses were analyzed, Wald $\chi2 = 8.14$, $p = .004$, suggesting that children increasingly discerned subtle yet biased messages in language, leading to a heightened tendency to associate natural ability with the reference group as they age. However, this interpretation is tentative as the main effect of age was not observed in the weighted responses ($B = 0.08$, $SE = 0.07$, $p = .276$).

Study 2*.* As in Study 1, Study 2 also found a significant main effect of condition on participants' ability beliefs in the forced-choice responses (Wald $\chi2 = 22.63$, $p < .001$), suggesting that children in the Equal Gesture condition were less likely to associate greater natural ability with the reference group than children in the Unequal Gesture condition. We also found a significant three-way interaction among condition, gender, and age (Wald $\chi2 = 5.84$, $p = .016$). When analyzing the weighted responses, we found a similar main effect of condition ($B = -0.91$, $SE = 0.15$, $p < .001$), a main effect of gender ($B = -0.31$, $SE = 0.15$, $p = .045$), and a parallel significant three-way interaction among condition, gender, and age

(*B* = -0.57, *SE* = 0.26, *p* = .031). To further investigate this three-way interaction, we performed the same mixed effects logistic regression model separately for each gender group. There was a significant main effect of condition for both boys (Wald $\chi 2 = 10.80$, $p = .001$) and girls (Wald $\chi 2 = 11.64$, $p < .001$), suggesting that both boys and girls were less likely to infer gender stereotypes when the SCS was accompanied by the *equal* gesture compared to the *unequal* gesture. However, we found a significant interaction between age and condition in girls' responses (age*condition, Wald $\chi 2 = 8.41$, $p = .004$), but not in boys' responses (age*condition, Wald $\chi 2 = 0.16$, $p = .693$). With age, girls in the Unequal Gesture condition became more likely to favor the reference group, whereas girls in the Equal Gesture condition became less likely to do so.

Supplemental Analyses (exploratory; not pre-registered)

Supplemental analyses were performed to explore three questions. First, how children's ability and effort judgments related to each other in the present studies. Second, whether the condition effect was consistent across the two different question types (ability vs. effort). Third, whether demographic variables (race/ethnicity, and SES) moderated the main effect of condition.

The relation between ability and effort judgements. In order to assess the relation between children's ability and effort judgments, we separated each participant's responses by question type (ability questions vs. effort questions). An ability score and an effort score for each child was created by recording the number of relevant trials on which children linked either natural ability or more required efforts to the reference group. For each trial, a selection of the reference group was coded as 1, whereas a selection of the variant group was coded as 0. The Pearson correlation analyses revealed a significant negative relation between children's ability and effort judgments in both Study 1 (r = -0.19, p = .014) and Study 2 (r = -0.18, p = .025). In line with past literature, these results suggest that children tend to reason that natural ability and effort are negatively related.

Question type. To explore whether the main effect of condition on children's stereotype endorsement was moderated by question type (ability vs. effort), we added question type as an additional predictor, as well as its interaction with other fixed factors, to the original mixed-effect logistic regression on each child's forced-choice responses. In addition, two-sample t-tests were performed to examine the effect of condition on children's ability and effort judgments separately.

Study 1. In the mixed-effect logistic regression model, a significant main effect of condition was still found (Wald χ^2 = 11.37, p = .001). The condition effect was *not* moderated by question type (condition*question type, Wald χ^2 = 2.48, p = .115), suggesting that the *equal* gesture's effect in mitigating children's stereotype acquisition was consistent in children's ability and effort judgments.

Results from two-sample t-tests revealed similar findings. When asked the ability questions, children in the *Equal* Gesture Condition (M = 0.38, SD = 0.40) were significantly less likely to choose the reference group compared to children in the No Gesture Condition (M = 0.59, SD = 0.39), t(158) = -3.49, p < .001. When asked the effort questions, children in the *Equal* Gesture condition (M = 0.48, SD = 0.41) tended to choose the reference group as requiring more effort than children in the No Gesture condition (M = 0.37, SD = 0.45), t(157) = 1.66, p = .099.

Study 2. The mixed-effect logistic regression model revealed a significant main effect of condition (Wald $\chi^2 = 21.70$, $p < .001$), which was not moderated by the type of questions asked (condition*question type, Wald $\chi^2 = 0.20$, $p = .651$).

When asked the ability questions, children in the *Equal* Gesture condition ($M = 0.43$, $SD = 0.43$) were significantly less likely to attribute greater natural ability to the reference group than children in the *Unequal* Gesture condition ($M = 0.69$, $SD = 0.42$), $t(158) = -3.93$, $p < .001$. When asked the effort questions, children in the *Equal* Gesture condition ($M = 0.47$, $SD = 0.41$) were significantly more likely to attribute greater effort to the reference group than children in the *Unequal* Gesture condition ($M = 0.24$, $SD = 0.36$), $t(155) = 3.72$, $p < .001$.

Overall, these findings suggest that the mitigating effects of the *equal* gesture on children's stereotype endorsement are robust regardless of which type of questions children received.

Race/Ethnicity. We first examined whether the mitigating effect of the *equal* gesture on children's stereotype inferences varied by participants' racial backgrounds. Because White participants made up approximately half of our sample ($N_{White} = 153$) among families who reported their racial information ($N = 296$), we combined participants from minority racial/ethnic groups into one group (children of color) to contrast with White children in our analysis. We conducted a mixed-effects logistic regression model with race/ethnicity (white children vs. children of color), condition (Study 1: no gesture vs. *equal* gesture; Study 2: *unequal* gesture vs. *equal* gesture), plus the interaction between race and condition as fixed effects, and a random intercept of participants.

Study 1. The results of Study 1 suggested that children's stereotype endorsement was mitigated by the presence of the *equal* gesture (main effect of condition, Wald $\chi2 = 11.25$, $p < .001$). Race did not moderate the main effect of condition on children's stereotypes (condition*race, Wald $\chi2 = 0.001$, $p = .972$). In other words, when the biased language was accompanied by the *equal* gesture, both White children ($M_{No\ Gesture} = 0.56$ vs. $M_{Equal\ Gesture} = 0.39$) and children of color ($M_{No\ Gesture} = 0.67$ vs. $M_{Equal\ Gesture} = 0.51$) were less likely to endorse a stereotype favoring the reference group's natural ability.

Study 2. We found a main effect of condition indicating that the *equal* gesture mitigated children's stereotype learning (main effect of condition, Wald $\chi2 = 23.20$, $p < .001$). Children's own racial background did not moderate this effect (condition*race, Wald $\chi2 = 0.91$, $p = .341$). Both White children ($M_{Unequal\ Gesture}$

= 0.77 vs. $M_{Equal\ Gesture}$ = 0.49) and children of color ($M_{Unequal\ Gesture}$ = 0.70 vs. $M_{Equal\ Gesture}$ = 0.49) were equally likely to benefit from the presence of the *equal* gesture and endorse egalitarian beliefs.

Socioeconomic Status (SES). Does the effect of the *equal* gesture on children's stereotypical beliefs differ for children from low- versus high-SES backgrounds? To probe the potential moderating effects of SES on condition, we created a composite SES measure by considering both caregiver(s) education level and household income. We first converted caregiver(s) education level to years of education and standardized it alongside each family's annual household income. These two standardized scores were then averaged to form a composite SES variable. We used this variable to replace the race variable in the same mixed-effects logistic regression model described above.

Study 1. A significant main effect of condition on children's ability inferences (Wald $\chi2 = 10.25$, $p = .001$) was found. SES did not moderate the main effect of condition (SES*condition, Wald $\chi2 = 1.07$, $p = .301$). Children from both high-SES (above median) and low-SES (below median) families showed a decrease in their stereotype endorsement when the *equal* gesture was present (High-SES children: $M_{No\ Gesture}$= 0.62 vs. $M_{Equal\ Gesture}$ = 0.44; Low-SES children: $M_{No\ Gesture}$= 0.60 vs. $M_{EqualGesture}$ = 0.45).

Study 2. The analysis revealed a main effect of condition (Wald $\chi2 = 20.94$, $p < .001$). Children's SES backgrounds did not significantly moderate this effect in children's stereotypical inferences (SES*condition, Wald $\chi2 = 0.17$, $p = .678$). Children who saw the *equal* gesture were less likely to endorse a stereotype favoring the reference group compared to children in the Unequal Gesture condition, regardless of their SES backgrounds (High-SES children: $M_{Unequal\ Gesture}$= 0.69 vs. $M_{Equal\ Gesture}$ = 0.54; Low-SES children: $M_{Unequal\ Gesture}$= 0.77 vs. $M_{Equal\ Gesture}$ = 0.47).

Final Comments.

Are the effects we found specific to the *equal* gesture? The most important aspect of the gesture we used in our study is that it must capture the notion of equality. We suspect that what is essential for the *equal* gesture to be effective in mitigating the effect of the SCS is that both hands be positioned on the vertical axis at the same height (and at different heights for the *unequal* gesture; the vertical plane is perceived as signaling competence, 6). However, the shape of the hand seems less crucial to the equality meaning than its placement. We speculate that a variety of handshapes could be effectively used in the *equal* gesture (e.g., 2 extended fingers are likely to work just as well as the 4 extended fingers we used).

We stress that our study focuses on co-speech gestures, not emblems. Emblems (e.g., the thumbs-up gesture indicating approval or the “OK” gesture) are gestures that can stand on their own and be produced without speech. They are culturally specific and have standards of form. For example, if you extend your pinky rather than your thumb, you are not producing a thumbs-up gesture—in the U.S., a pinky-up gesture has no meaning; in China, it indicates disapproval rather than approval. In contrast, the *equal* and *unequal* gestures used in our study do not convey meaning on their own but instead carry meaning only when integrated with the speech context in which they occur.

Other co-speech gestures are likely to work to mitigate the biasing effect of the SCS, as long as the spatial alignment requirement that both hands be parallel on the vertical plane is met. Varying the gesture used to convey equality would be an interesting question for future work.

SI References

1. F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* **39**, 175–191 (2007).
2. E. K. Chestnut, M. Y. Zhang, E. M. Markman, “Just as good”: Learning gender stereotypes from attempts to counteract them. *Developmental Psychology* **57**, 114–125 (2021).
3. K. Muenks, D. B. Miele, Students’ Thinking About Effort and Ability: The Role of Developmental, Contextual, and Individual Difference Factors. *Review of Educational Research* **87**, 707–735 (2017).
4. M. Klem, et al., Sentence repetition is a measure of children’s language skills rather than working memory limitations. *Developmental Science* **18**, 146–154 (2015).
5. M. Komeili, C. R. Marshall, Sentence repetition as a measure of morphosyntax in monolingual and bilingual children. *Clinical Linguistics & Phonetics* **27**, 152–162 (2013).
6. A. Lakshmi, “Gesture and stereotype content: Inquiries on perception, activation, and spontaneous production,” Doctoral dissertation, The University of Chicago, Chicago, IL. (2024).