How does Language Bias Memory for Gender Stereotypes?

Meghan Kalvey
Department of Psychology
New York University
meghan.kalvey@nyu.edu

Cynthia Boo
Department of Psychology
New York University
cynthia.boo@nyu.edu

Marjorie Rhodes
Department of Psychology
New York University
marjorie.rhodes@nyu.edu

Abstract

Combating stereotypes ("girls do ballet") may not be as simple as conveying counter-stereotypical messages ("boys do ballet"). For example, hearing generic language ("girls/boys do X") predicted greater stereotyping among young children than hearing specific language (that girl/boy does X"), even when parents explicitly conveyed counter-stereotypical messages (Benitez et al., 2025). Why might counter-stereotypical messages delivered generically backfire, reinforcing the very stereotypes they aim to combat? We propose that generic language activates category-level representations stored in memory, bringing to mind stereotypical information about the implicated categories. 183 adults and 104 4- to 6-year-olds recalled the language (generic vs. specific) and message (stereotypical vs. counter-stereotypical) heard in 12 statements. Adults, but not children, systematically misremembered counter-stereotypical messages as stereotypical ones when language was generic. The task was difficult for children, perhaps explaining their recall patterns and high exclusion rate (48% of 200 children were excluded). Follow-up work probing recognition memory is underway.

Keywords: social cognition; stereotypes; generics; memory; development; mental representations

Introduction

Stereotypes—such as "boys are good at math" and "girls are nurturing"—are widely shared, overgeneralized beliefs about social groups that are applied to individuals based on group membership, regardless of personal variation (Marx & Ko, 2019; Vinacke, 1957). They are often learned through cultural transmission, particularly linguistic communication, rather than through direct experience (Kashima et al., 2013; Robles & Kurylo, 2017; Thompson et al., 2000). Stereotypes develop by age four (Foster-Hanson & Rhodes, 2023; Miller et al., 2009; Sherman et al., 2016) and have well-documented negative effects in childhood and adulthood at both personal and societal levels (Heilman, 2012; Kellow & Jones, 2008; Meeussen et al., 2022; Wang et al., 2025).

Linguistic interventions aimed at combating stereotypes often involve communicating counter-stereotypical messages (e.g., "girls play superhero [too]"; Fraser et al., 2023; Mun et al., 2023). Notably, counter-stereotypical messages, just like stereotypical ones, often reference categories using generic language (hereon, "generics"). Generics (statements such as "*girls* bake cookies" and "*boys* play princess") refer to categories as abstract wholes, rather than to specific individuals or subsets (hereon, "specifics"; statements such as "*that girl* bakes cookies" or "*those boys* play princess").

In a naturalistic task where parents discussed a picture book depicting a boy or a girl performing various activities (e.g., a boy putting on makeup), Benitez et al. (2025) found that parents often referenced gender categories generically. Further, parents' use of generics predicted greater gender stereotyping in 3- to 5-year-olds, even when parental speech explicitly conveyed counter-stereotypical messages. In other words, counter-stereotypical messages when delivered as generics appear to backfire, reinforcing the very stereotypes they aim to combat. Here, we propose that generics backfire in part because of how such language is processed and stored in memory. In particular, we propose that generics—regardless of the message stereotypicality—activate category-level representations that bias memory toward stereotypical information. As a result, counter-stereotypical messages delivered as generics may be distorted in memory processes, and thereby misremembered as generic stereotypical information instead.

Both children and adults preferentially encode and recall generic statements, and are more likely to misremember non-generic statements as generic ones than the reverse (Gelman et al., 2019; Gelman & Raman, 2007; Sutherland et al., 2015). For example, Leslie & Gelman (2012) showed preschoolers and college students pictures of animals accompanied by spoken statements that were either generic ("spiders shed their skin") or non-generic (e.g., "some hippos like to swim"). Then, after a 4-minute distractor task, children and adults were asked to recall what they heard when they saw each picture. Overall, authors found a tendency in children and adults to recall generics more accurately than non-generics, and for non-generics to be mis-recalled as generics more than the reverse. This pattern also holds across languages where sentence length is equal for generic and non-generic statements (Gelman et al., 2016). Separately, there is evidence that stereotypical information is often better remembered than counter-stereotypical information (Fyock & Stangor, 1994). However, these two literatures have largely been studied in isolation.

Here, we test whether interactions between langauge (generic vs. specific) and message (stereotypical vs. counter-stereotypical) might lead to systematic memory distortions that perpetuate stereotypes. We considered that two processes could contribute to such distortions, reinforcing prior

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

stereotypical beliefs in memory rather than storing the counter-stereotypical message that was explicitly said. First, if people hear a generic counter-stereotypical statement (such as "boys wear nail polish") they could bring to mind prior stereotypical beleifs and reject what was explicitly said ("no, boys don't wear nail polish"). Second, people could bring to mind alternate categories that are stereotypically associated with the predicate, also fitting their prior beliefs ("but girls do ballet!"). In other words, because generic counter-stereotypical statements do not fit with listener's prior category-level knowledge and beleifs (i.e., existing mental representations of what "boys" and, by extension, "girls" as distinct categories are like), listeners could respond by bringing to mind different predicate information ("boys *don't* wear nail polish") or different categories that are more consistent with the mentioned predicate ("*girls* wear nail polish"). The second process would predict a particular pattern of memory distortions—listeners would misremember generic counter-stereotypical statements as generic stereotypical ones by switching the category during recall to stereotypically fit the mentioned predicate. This process should be restricted to the cases where counter-stereotypical messages are delivered generically, because non-generic statements describing particular individuals displaying counter-stereotypical behaviors (such as "*that boy* wears nail polish") make listeners less likely to bring to mind category-level representations (Moty & Rhodes, 2021)

Given prior work on memory for generics (Gelman et al., 2019; Gelman & Raman, 2007; Gelman et al., 2016; Leslie & Gelman, 2012; Sutherland et al., 2015), we hypothesized that people would show systematic memory biases, incorrectly recalling specific statements as generic ones more than the reverse. In addition, given schema effects on memory (Fyock & Stangor, 1994), we hypothesized that people would recall counter-stereotypical messages as stereotypical ones more than the reverse. Critically, we also predicted an interaction in memory patterns, such that people would be particularly likely to mis-recall generic counter-stereotypical statements as generic stereotypical ones.

Present Study

In Study 1, we examined recall accuracy and errors in adults. In Study 2, we examined the same in 4- to 6-year-olds.

We aimed to detect Cohen's $d = .50$ (OR ≈ 2.0) with 95% power at $\alpha = .05$, consistent with prior studies reporting medium to large effects (Gelman et al., 2019; Gelman & Raman, 2007; Leslie & Gelman, 2012; Sutherland et al., 2015). An a priori power analysis for a 2 (language: generic vs. specific) × 2 (message: stereotypical vs. counter-stereotypical) factorial design based on generalized linear mixed-effects modeling (GLMM) with a logit link function for binary recall accuracy (correct vs. incorrect) indicated a target sample size of 180 adults (Study 1) and 180 children (Study 2). We oversampled and recruited 200 participants per study to account for attrition, technical issues, and exclusion due to inadequate participation.

Participation was considered inadequate if a participant recalled nothing (i.e., silence during trials or providing off-task responses) or if their recalled statements were incomplete (i.e., stating a feature, such as "basketball", without referencing a category) for more than 4 out of the 14 items (12 gendered and 2 control). Responses were considered adequate if participants referenced categories in their responses, such that we could identify the language and message that was recalled.

Hypotheses, study procedures, exclusion criteria, and analysis plan were pre-registered (osf.io/5f2wz) and the study methods were approved by the New York University Institutional Review Board (IRB-FY2025-10046).

Study 1: Adults

Methods

Adult participants were recruited via Prolific and participated on a survey hosted on Qualtrics. All participants were fluent English speakers, primarily located in the United States, and provided informed consent. Data were collected in April 2025, asynchronously without real-time researcher supervision. Participants were compensated $4.25.

Of the 200 recruited adults, 8.5% were excluded using the aforementioned pre-registered criteria. Only data from included adult participants were analyzed and are reported below.

Participants. Adults ($N = 183$; 91 women, 90 men, 2 non-binary; $M_{age} = 39.19$, $SD = 12.21$, range: 19–69) self-identified their race/ethnicity as: White (63.39%), Black (23.49%), Hispanic (4.92%), Asian (4.92%), Multiracial (2.73%), and Native (0.55%). Adults self-reported their level of education as: High School diploma/GED (22.95%), Associate's degree (10.38%), Bachelor's degree (44.26%), Master's degree (19.67%), and Doctoral degree (2.19%). One adult did not report their level of education.

Procedure. The task was a modified version of the popular children's game, *Telephone*. Participants saw pictures and heard statements about each picture from a fictional gender-neutral character named Jamie. Then, participants were asked to tell the narrator exactly what Jamie told them about each picture.

In the training phase (*Figure 1*), participants were familiarized with the task. Modified versions of the stimuli from Gelman and Raman (2007) were used here.

In the learning phase (*Figure 2*), participants saw 14 pictures accompanied by auditory statements, just like the stimuli in the training phase. Two test items were controls, and 12 referenced gender categories. All normed in prior studies (Signorella & Liben, 1984; Sullivan et al., 2022; Sullivan et al., 2018), six items addressed girl-stereotyped activities (e.g., "making jewelry", "wearing nail polish") and six addressed boy-stereotyped ones (e.g., "playing with tools," "riding motorcycles").

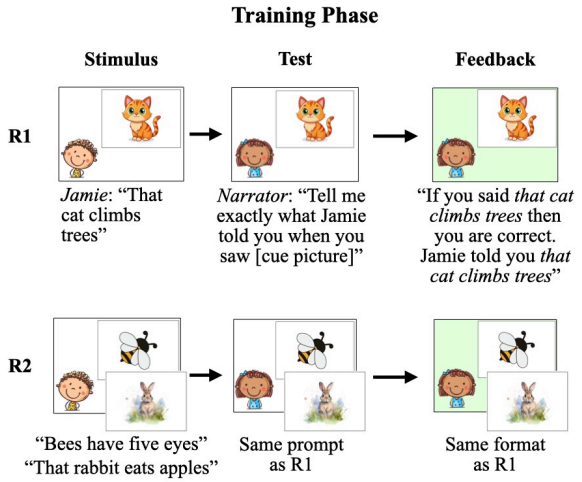

Figure 1: Training phase setup. In round 1 (R1), participants recalled one statement and received feedback. In round 2 (R2), participants recalled two statements delivered consecutively, followed by feedback.

All participants saw the same 14 pictures and heard the same 2 control statements. But the 12 gendered statements varied across participants based on the block to which they were assigned. Blocks were created to counterbalance the four conditions formed by crossing language and message: generic stereotypical (e.g., "boys play basketball"), generic counter-stereotypical (e.g., "girls play basketball"), specific stereotypical (e.g., "that boy plays basketball"), and specific counter-stereotypical (e.g., "that girl plays basketball"). Each item was presented in one of the four conditions such that each participant saw 3 items per condition. An equal number of participants were randomly assigned to each block ($\chi^2(3) = .19$, $p = .98$), and item order was randomized to mitigate serial position effects on recall (Forrin & MacLeod, 2016; Mulligan & Lozito, 2007).

Learning Phase

Item	Block 1	Block 2	Block 3	Block 4
	generic stereotypical *"girls play princess"*	generic counter-stereotypical *"boys play princess"*	specific stereotypical *"that girl plays princess"*	specific counter-stereotypical *"that boy plays*
	specific counter-stereotypical *"that girl plays superhero"*	generic stereotypical *"boys play superhero"*	generic counter-stereotypical *"girls play superhero"*	specific stereotypical *"that boy plays superhero"*
3 items per condition in each block				
	**control**: "Jamie roller skates on Sundays"			
	**control**: "Jamie had chocolate for dinner"			

Figure 2: Learning phase setup. Here we illustrate two example items, addressing a girl-stereotyped (plays princess) and a boy-stereotyped activity (plays superhero). Blocks show how conditions were counterbalanced for each item. Participants were randomly assigned to a block; each participant heard statements in all conditions varied by item. The two control items were the same across blocks.

After hearing the 14 statements, participants watched 4-minutes of *Bluey minisodes* as a distractor task to prevent rehearsal. The distractor duration matched prior work (Gelman et al., 2019; Gelman & Raman, 2007; Leslie & Gelman, 2012; Sutherland et al., 2015). We ensured that visual and linguistic content in these episodes ("Hungry" and "Drums") were unrelated to the stimuli used in this study.

Finally, in the test phase, participants saw the same 14 pictures from the learning phase, presented in a new randomized order. Using the same test prompt from the training phase, adults were asked type their recall responses in a textbox below each picture.

At the very end, participants completed a brief demographic questionnaire.

Data Processing and Analysis. Data processing and analyses were conducted in R (version 4.4.1).

We first cleaned the data and excluded participants using pre-registered criteria. In this process, we identified a programming error: blocks 1 and 3 contained 3 items per condition (as intended) but blocks 2 and 4 contained an unequal distribution of conditions. To evaluate the impact of this error on our findings, we tested whether block assignment (blocks 1 & 3 vs. blocks 2 & 4) predicted recall accuracy as we expected significant variation in recall accuracy by condition. Reassuringly, recall accuracy was not predicted by block assignment (b = -0.25, 95% CI [-0.69, 0.20], OR = 0.78, p = .28), therefore, we used all data without additional processing.

Adults' typed responses were imported from Qualtrics and annotated using string matching in R. Referenced categories were identified within the first two words of their responses. The message was annotated as "stereotypical" if participants recalled masculine labels (e.g., boy, man, guy, he, etc.) for boy-stereotyped activities or feminine labels (e.g., girl, woman, lady, she, etc.) for girl-stereotyped activities. Similarly, the message was annotated as "counter-stereotypical" if participants recalled masculine labels for girl-stereotyped activities and feminine labels for boy-stereotyped activities. If participants recalled non-gendered labels (e.g., person, kid, they, etc.) or recalled both gendered labels neutrally (e.g., "girls and boys do X"), then the message was annotated as "neutral".

The language was annotated as "generic_strict" if participants recalled labels using bare plurals (e.g., boys, girls, men, women, guys, ladies, etc.); as "specific_strict" if they paired labels with singular demonstratives (e.g., that/this girl does X); as "specific_lenient" if they paired labels with plural determiners (e.g., these/those boys do X), quantifiers (e.g., some/most/all girls do X), the article "the" (e.g., the boy does X), or used gendered pronouns (she/he does X); as "ambiguous" if they recalled labels as bare nouns (e.g., girl/boy does X), used pronoun "they", or paired label with the article "a" (e.g., a girl does X).

If participants did not reference a gender or neutral category, then the message and language were annotated as "NA" for those items.

Recall accuracy was scored by comparing the language and message in participants' recalled responses to corresponding items at learning. Language accuracy was scored as "correct" if the recalled language matched that of the stimulus, and "incorrect" otherwise. Similarly, message accuracy was scored as "correct" if the recalled message matched that of the stimulus, and "incorrect" otherwise. An item was scored as "correct" only when both the language and message were recalled correctly.

Participant responses annotated as "neutral", "ambiguous", or "NA" were treated as "incorrect" for analyses. As per the exclusion criteria, the included participants had no more than 4 "NA" annotated responses.

Control items were used to verify comprehension and attention; recall accuracy and error analyses were conducted on the 12 gendered items only.

Results

Recall accuracy and errors were analyzed using GLMM ("lme4" package in R). Models controlled for participant- and item-level variability. Additionally, we conducted four separate Chi-square post-hoc pairwise comparisons with a Bonferroni correction (one per stimuli condition) to compare three types of error conversions:

(1) message: stereotypical misremembered as counter-stereotypical, or vice versa
(2) language: generics misremembered as specifics, or vice versa
(3) both: message and language misremembered (e.g., generic stereotypical items misremembered as specific counter-stereotypical ones)

Recall accuracy for language was scored using strict and lenient approaches (i.e., recalled responses annotated as "specific_strict" and "specific_lenient"). Model outputs for both strict and lenient approaches yielded equivalent results, so we only report results for strictly coded responses.

Recall Accuracy. Overall, adults recalled 48.81% of items accurately (both the language and message were recalled accurately for these items). Adults recalled the language accurately for 55.50% of items, and message accurately for 79.37% of items.

Accuracy patterns replicated prior work but did not reveal an interaction (*Figure 3*). Language and message independently predicted recall accuracy as generics ($M_{likelihood}$ = 0.66) were recalled more accurately than specifics ($M_{likelihood}$ = 0.31; b = 1.82, 95% CI [1.52, 2.13], OR = 6.23, p < .001), and stereotypical messages ($M_{likelihood}$ = 0.52) were recalled more accurately than counter-stereotypical ones ($M_{likelihood}$ = 0.45; b = 0.39, 95% CI [0.09, 0.69], OR = 1.48, p < .01). Language and message did not interactively predict recall accuracy (b = 0.08, 95% CI [-0.34, 0.50], OR = 1.08, p = .71).

Recall Errors. Error patterns replicated prior work and revealed the hypothesized interaction (*Figure 4*).

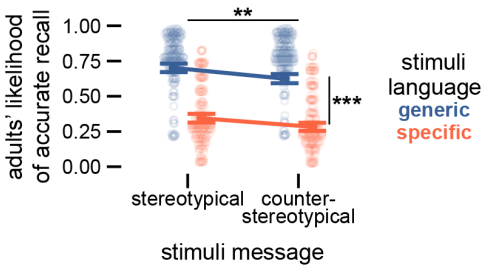

Figure 3: Recall accuracy in adults. Each datapoint represents one adult's likelihood of accurate recall as a function of the stimuli condition. Error bars are 95% CIs. **p < .01; ***p < .001.

Specifics were misremembered as generics more than the reverse ($\chi^2(1)$ = 234.48, p < .001), and counter-stereotypical messages were misremembered as stereotypical ones more than the reverse ($\chi^2(1)$ = 38.91, p < .001). Critically, when language was generic, adults were most likely to misremember the message (counter-stereotypical as stereotypical, and vice versa; contrast ps < .01). Whereas, when language was specific, adults were most likely to misremember the language itself (specific as generic; contrast ps < .001).

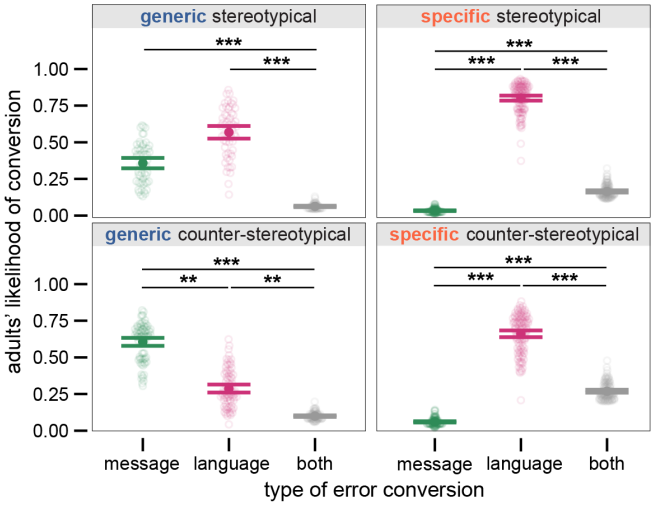

Figure 4: Recall errors in adults. Each datapoint represents one adult's likelihood for each type of error conversion as a function of stimuli condition (represented by panels). Error bars are 95% CIs. **p < .01; ***p < .001.

Discussion

Consistent with prior work, findings in adults show generic bias and stereotype bias in memory. Additionally, we found the hypothesized error pattern: adults systematically converted the message when language was generic but not when it was specific, particularly resulting in generic counter-stereotypical items being misremembered as generic stereotypical ones. These error patterns suggest that counter-stereotypical generics backfire by reinforcing stereotypes stored in memory.

Study 2: Children

Methods

Children were recruited via our lab's online platform to conduct unmoderated remote research with children and their families, PANDA (discoveriesinaction.org). The survey was hosted on Qualtrics. Participants were English speakers primarily in the United States. Parents provided informed consent and children provided assent. Data were collected in May 2025. Families were compensated with $10 Amazon gift cards.

Of the 200 recruited children, 48% were excluded using pre-registered criteria. This high exclusion rate was unexpected. This task was too difficult for young children as child age predicted the number of adequate recall responses in the recruited sample (b = 2.11, 95% CI [1.84, 2.37], OR = 8.23, p < .001).[1] As a result of this exclusion rate, the child sample here is under-powered to detect any language × message interactions. Analyses from the included sample only are reported below.

Participants. Four- to 6-year-olds' (N = 104, 62 girls, 41 boys, 1 non-binary; M_{age} = 5.01, SD = 0.62, range: 4–5.98) race/ethnicity was reported by their parents as: White (58.65%), Black (4.81%), Hispanic (5.77%), Asian (7.69%), MENA (3.85%), and Multiracial (19.23%). And equal number of boys and girls were recruited, but unexpectedly, more girls than boys surpassed the exclusion criteria ($\chi^2(1)$ = 4.28, p < .05).[2]

Procedure. The study task and materials were identical to Study 1, except that children provided spoken responses that were recorded on PANDA via their computer's webcam and microphone.

Data Processing and Analysis. Children's spoken responses were transcribed using Datavyu (a video annotation software; datavyu.org). To determine the reliability of transcripts, a second coder transcribed a random 25% of files. Agreement between coders was substantial (mean κ = .76, range: .31–1). Files with low κ typically reflected speech intelligibility due to audio quality or the child's manner of speaking.

Transcribed responses were imported from Datavyu into R. Data processing and analysis procedure in R were identical to Study 1, except that the referenced categories in children's spoken responses were identified using a different technique. While adults in Study 1 consistently included gender categories in the first two words of their responses, children's utterances varied widely. Many children used fillers (e.g., "um, um, girls do ballet"), stuttered (e.g., "bo…boys ride motorcycles"), self-corrected (e.g., "boys…girls play basketball"), or paired their responses with off-task conversation (e.g., "Mom I'm hungry…girls play with tools"). Therefore, to identify the categories referenced by children at recall, we searched utterances backwards (starting with the last word produced) using string matching in R. The message and language were then annotated using the same rules described in Study 1.

The programming error related to unequal distribution of conditions across blocks did not predict recall accuracy in children either (b = 0.20, 95% CI [−0.21, 0.60], OR = 1.22, p = .34).

Results

Recall Accuracy. Overall, children recalled 45.19% of items accurately. Children recalled the language accurately for 49.76% of items, and message accurately for 75.40% of items.

Accuracy patterns replicated some prior work and did not reveal an interaction (*Figure 5*). Language alone predicted recall accuracy as generics ($M_{likelihood}$ = 0.55) were recalled more accurately than specifics ($M_{likelihood}$ = 0.35; b = 0.74, 95% CI [0.39, 1.09], OR = 2.09, p < .001). Stereotypical ($M_{likelihood}$ = 0.46) and counter-stereotypical messages ($M_{likelihood}$ = 0.44) were recalled equally accurately (b = -0.06, 95% CI [-0.42, 0.29], OR = 1.48, p = .73). Language and message did not interactively predict recall accuracy (b = 0.296, 95% CI [-0.20, 0.796], OR = 1.08, p = 0.24).

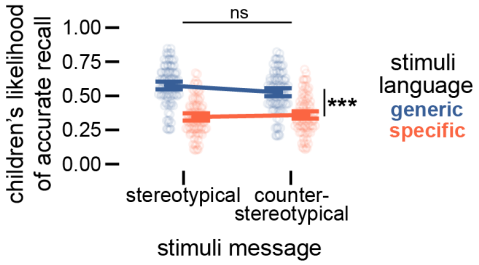

Figure 5: Recall accuracy in children. Each datapoint represents one child's likelihood of accurate recall as a function of stimuli condition. Error bars are 95% CIs. ***p < .001.

Recall Errors. Error patterns replicated prior work but did not reveal the hypothesized interaction (*Figure 6*). Specifics were misremembered as generics more than the reverse ($\chi^2(1)$ = 43.94, p < .001), and counter-stereotypical messages were misremembered as stereotypical ones more than the reverse ($\chi^2(1)$ = 7.07, p < .01). However, for all four stimuli conditions, children were most likely to misremember the language (generics as specifics, and vice versa; contrast ps < .01).

[1]Language and message did not systematically predict the number of "NA" annotated responses in children (bs < 0.16, ps > .12); children's tendency to provide inadequate responses was random and not determined by the stimuli condition.

[2]Girls did not show superior memory performance as recall accuracy was not predicted by children's reported gender (b = −0.16, 95% CI [−0.56, 0.23], OR = 0.85, p = .42; this analysis excluded non-binary children). At present, we do not have an explanation for why more girls surpassed the exclusion criteria.

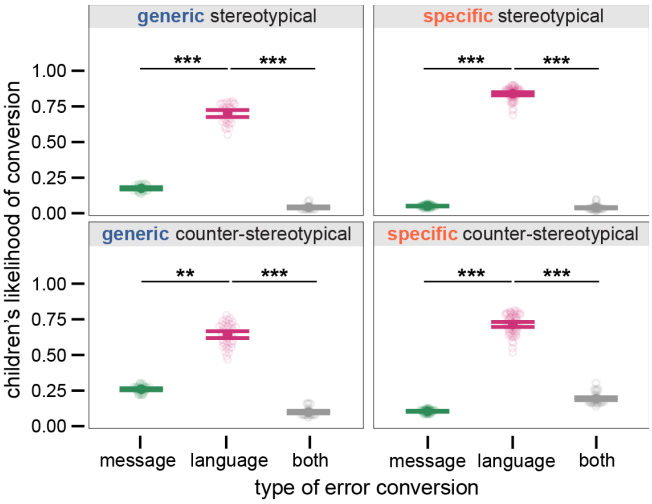

Figure 6: Recall errors in children. Each datapoint represents one child's likelihood for each type of error conversion as a function of stimuli condition (represented by panels). Error bars are 95% CIs. $^{**}p < .01$; $^{***}p < .001$.

Discussion

Partially consistent with prior work, findings in children show generic bias, but not stereotype bias, in memory. Moreover, we did not find the hypothesized error pattern as children mostly converted the language at recall (generic to specific, and vice versa).

General Discussion

Adults (but not children) systematically misremembered counter-stereotypical messages as stereotypical ones when language was generic but not when specific. Consistent with our initial proposal, this pattern suggests that among adults hearing generic counter-stereotypical statements, in particular, reinforces stereotypes in memory.

Counter-stereotypical generics may activate stereotypes in memory by two processes. First, generics could activate conflicting background knowledge, expressed by negating the predicate (e.g., upon hearing "boys make jewelry", listeners bring to mind, "boys *don't* make jewelry"). Second, generics could bring to mind alternate categories that align with the mentioned predicate (e.g., "girls make jewelry"). In the present study, all conversions from counter-stereotypical to stereotypical messages reflected category switches (e.g., adults and children misremember "boys make jewelry" as "girls make jewelry"), consistent with the second process. Thus, one possible explanation for the differences in recall patterns among adults and children, if found to be robust, is that children do not bring to mind alternate categories in this task. This could be due to high task demands or it could reflect real developmental change in the tendency to bring to mind alternate categories in language processing. Identifying whether these findings reflect real developmental change is the goal of our future work. Furthermore, this tendency in adults to switch categories instead of negating the predicate might result from the present task design, which did not include any negated sentences at learning and provided a visual cue to the predicate at test. Hence, future work will also attempt to identify if memory errors that involve negating the predicate arise under other circumstances.

Conclusion

Despite egalitarian intentions, counter-stereotypical messages backfire when delivered as generics, reinforcing the very stereotypes they aim to combat. Understanding cognitive processes that drive subtle effects—such as how generics could activate stereotypes in memory—is crucial to effectively combat the development of harmful stereotypes.

Acknowledgements

We thank Julia Johnson for pre-processing participant data from PANDA, undergraduate research assistants Lauren Caron, Neela Rao, Alan Zhang, and Yijun Zhou for transcribing children's responses using Datavyu, and members of the Conceptual Development and Social Cognition Lab at NYU for feedback on this paper. We also thank adults on Prolific and participating families on PANDA for their time.

ChatGPT was used to debug some R code, and all input provided by ChatGPT was reviewed prior to use.

References

Benitez, J., Foster-Hanson, E., & Rhodes, M. (2025). Generic references to gender predict essentialism and stereotyping even when they express counter-stereotypic ideas. *J Exp Psychol Gen*. https://doi.org/10.1037/xge0001871

Forrin, N. D., & MacLeod, C. M. (2016). Order Information is Used to Guide Recall of Long Lists: Further Evidence for the Item-Order Account [Article]. *Canadian Journal of Experimental Psychology*, *70*(2), 125-138. https://doi.org/10.1037/cep0000088

Foster-Hanson, E., & Rhodes, M. (2023). Stereotypes as prototypes in children's gender concepts. *Developmental Science*, *26*(4), e13345. https://doi.org/https://doi.org/10.1111/desc.13345

Fraser, K. C., Kiritchenko, S., Nejadgholi, I., & Kerkhof, A. (2023). What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023),

Fyock, J., & Stangor, C. (1994). The role of memory biases in stereotype maintenance. *British Journal of Social Psychology*, *33*(3), 331-343.

Gelman, S. A., Leslie, S.-J., Gelman, R., & Leslie, A. (2019). Do children recall numbers as generic? A strong test of the generics-as-default hypothesis. *Language Learning and Development*, *15*(3), 217-231.

Gelman, S. A., & Raman, L. (2007). This cat has nine lives? Children's memory for genericity in language. *Developmental Psychology*, *43*(5), 1256.

Gelman, S. A., Tapia, I. S., & Leslie, S.-J. (2016). Memory for generic and quantified sentences in Spanish-speaking children and adults. *Journal of Child Language*, *43*(6), 1231-1244.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113-135. https://doi.org/https://doi.org/10.1016/j.riob.2012.11.003

Kashima, Y., Lyons, A., & Clark, A. (2013). The maintenance of cultural stereotypes in the conversational retelling of narratives. *Asian Journal of Social Psychology*, *16*(1), 60-70.

Kellow, J. T., & Jones, B. D. (2008). The Effects of Stereotypes on the Achievement Gap: Reexamining the Academic Performance of African American High School Students. *Journal of Black Psychology*, *34*(1), 94-120. https://doi.org/10.1177/0095798407310537

Leslie, S. J., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults [Article]. *Cognitive Psychology*, *64*(3), 186-214. https://doi.org/10.1016/j.cogpsych.2011.12.001

Marx, D., & Ko, S. J. (2019). Stereotypes and Prejudice. In: Oxford University Press.

Meeussen, L., Van Rossum, A., Van Laar, C., & Derks, B. (2022). Gender stereotypes: What are they and how do they relate to social inequality? In *Solidarity and social justice in contemporary societies: An interdisciplinary approach to understanding inequalities* (pp. 79-86). Springer.

Miller, C. F., Lurye, L. E., Zosuls, K. M., & Ruble, D. N. (2009). Accessibility of Gender Stereotype Domains: Developmental and Gender Differences in Children. *Sex Roles*, *60*(11), 870-881. https://doi.org/10.1007/s11199-009-9584-x

Moty, K., & Rhodes, M. (2021). The Unintended Consequences of the Things We Say: What Generic Statements Communicate to Children About Unmentioned Categories [Article]. *Psychological Science*, *32*(2), 189-203. https://doi.org/10.1177/0956797620953132

Mulligan, N. W., & Lozito, J. P. (2007). Order information and free recall: Evaluating the item-order hypothesis. *Quarterly Journal of Experimental Psychology*, *60*(5), 732-751. https://doi.org/10.1080/17470210600785141

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. *arXiv preprint arXiv:2311.00161*.

Robles, J. S., & Kurylo, A. (2017). 'Let's have the men clean up': Interpersonally communicated stereotypes as a resource for resisting gender-role prescribed activities. *Discourse Studies*, *19*(6), 673-693. https://doi.org/10.1177/1461445617727184

Sherman, S. J., Sherman, J. W., Percy, E. J., & Soderberg, C. K. (2016). Stereotype development and formation.

Signorella, M. L., & Liben, L. S. (1984). Recall and reconstruction of gender-related pictures: Effects of attitude, task difficulty, and age. *Child Development*, 393-405.

Sullivan, J., Ciociolo, A., & Moss-Racusin, C. A. (2022). Establishing the content of gender stereotypes across development. *PloS one*, *17*(7), 9-14.

Sullivan, J., Moss-Racusin, C., Lopez, M., & Williams, K. (2018). Backlash against gender stereotype-violating preschool children. *PloS one*, *13*(4), 6-7.

Sutherland, S. L., Cimpian, A., Leslie, S. J., & Gelman, S. A. (2015). Memory errors reveal a bias to spontaneously generalize to categories. *Cognitive science*, *39*(5), 1021-1046.

Thompson, M. S., Judd, C. M., & Park, B. (2000). The Consequences of Communicating Social Stereotypes. *Journal of Experimental Social Psychology*, *36*(6), 567-599. https://doi.org/https://doi.org/10.1006/jesp.1999.1419

Vinacke, W. E. (1957). Stereotypes as social concepts. *The journal of social Psychology*, *46*(2), 229-243.

Wang, M. M., Cardarelli, A., Brenner, J., Leslie, S. J., & Rhodes, M. (2025). Maladaptive but malleable: Gender-science stereotypes emerge early but are modifiable by language. *Child Development*, *96*(2), 865-880. https://doi.org/10.1111/cdev.14213