Journal of Experimental Child Psychology 191 (2020) 104753

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Counterstereotyping can change children's thinking about boys' and girls' toy preferences

Rachel Ann King [a,b,*], Katharine E. Scott [a], Maggie P. Renno [a], Kristin Shutts [a]

[a] University of Wisconsin–Madison, Madison, WI 53706, USA
[b] Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 21 June 2019
Revised 16 October 2019
1 November 2019
Available online 13 December 2019

Keywords:
Children
Gender
Stereotypes
Generics
Social cognitive development
Intervention

ABSTRACT

Children think that peers prefer gender-stereotypical toys over gender-counterstereotypical toys. These beliefs can limit children's exploration of gender-counterstereotypical behaviors and prevent the development of broad skills and interests. The current research tested interventions to combat gender-based stereotyping about toys among children aged 4 to 7 years (N = 373). Across four experiments featuring seven different intervention versions, participants saw videos where a teacher provided counterstereotypical messages about toy preferences (e.g., "boys like dolls," "girls like trucks"). The phrasing of the messages (e.g., generic vs. demonstrative) and accompanying photographs (e.g., images of many children vs. one child) varied across experiments. In all intervention conditions, participants made more counterstereotypical (and fewer stereotypical) predictions about peers' toy preferences after viewing intervention videos; differences in the phrasing of the intervention message (e.g., "boys like dolls" vs. "this kid likes dolls") had little effect on participants' predictions. In Experiment 4, an intervention condition containing generic phrasing and gender noun labels (e.g., "boys like dolls") changed children's selection of toys for peers. This research provides promise for counterstereotyping as an impactful and easily implementable intervention strategy.

* Corresponding author at: Cornell University, 211 Uris Hall, Ithaca, NY, 14853, USA. Fax: +1 607 255 8433.
E-mail address: rak295@cornell.edu (R.A. King).

https://doi.org/10.1016/j.jecp.2019.104753
0022-0965/

Introduction

Children's environments are replete with messages that girls and boys like fundamentally different kinds of things. Storybooks (Berry & Wilkins, 2017; Gooden & Gooden, 2001), toy marketing (Blakemore & Centers, 2005; Murnen, Greenfield, Younger, & Boyd, 2016; Reich, Black, & Foliaki, 2018), television (Davis, 2003; Kahlenberg & Hein, 2010; Leaper, Breed, Hoffman, & Perlman, 2002), and movies (Arnold, Seidl, & Deloney, 2015; Streiff & Dundes, 2017) portray stereotypical representations of boys' and girls' interests. For example, the girls' section of the Disney Store's website features pink and purple colors, jewelry, and cosmetics, whereas the boys' section features red, black, and brown colors, building toys, weapons, and vehicles (Auster & Mansbach, 2012).

Children are also exposed to stereotypes in their own homes. Beginning in infancy, parents often fill boys' and girls' rooms with gender-stereotypical items: Boys' rooms typically display sports equipment, vehicles, and the color blue, whereas girls' rooms typically display jewelry, dolls, and the color pink (MacPhee & Prendergast, 2019; Pomerleau, Bolduc, Malcuit, & Cossette, 1990; Witt, 1997). Parents also tend to buy their children gender-stereotypical toys (Weisgram & Bruun, 2018) and reject their children's requests for counterstereotypical gifts (Etaugh & Liss, 1992; Robinson & Morris, 1986). Furthermore, parents—even those with egalitarian motives—often steer their boys toward stereotypically masculine activities and away from stereotypically feminine ones (Freeman, 2007; Halpern & Perry-Jenkins, 2016; Kane, 2006).

Children's toy and play stereotypes

Beginning early in life, children are aware of, and behave in accordance with, stereotypical messages in their environments (for complete reviews, see Martin & Ruble, 2004; Ruble, Martin, & Berenbaum, 2007). For example, young children associate stereotypically feminine items such as dolls and tea sets with girls but not with boys, and they associate stereotypically masculine items such as trucks and tools with boys but not with girls (e.g., Cherney & Dempsey, 2010; Freeman, 2007; Todd et al., 2018). In addition to being knowledgeable about gender stereotypes regarding toy preferences, young children act in alignment with widely held stereotypes. For example, when asked to choose toys for others, 3- to 5-year-olds avoid choosing trucks for girls (favoring dolls instead) and avoid choosing dolls for boys (favoring trucks instead) (Cowan & Hoffman, 1986; Eisenberg, Murray, & Hite, 1982). Children of this age also make similar decisions when choosing toys for themselves (Fabes, Martin, & Hanish, 2003; Halim, Ruble, Tamis-LeMonda, & Shrout, 2013; Zosuls et al., 2009). Furthermore, young children will exclude children whose gender does not match the gender commonly associated with the play activity at hand (Theimer, Killen, & Stangor, 2001).

Despite ample research documenting children's gender stereotyping, less is known about effective strategies for addressing early-emerging stereotypical beliefs. Yet, reducing children's stereotyping is important because such biases can constrain children's interest in, and access to, advantageous experiences. For example, block play, which is stereotypically associated with boys (Miller, 1987; Sherman, 1967), provides opportunities to hone spatial skills (Levine, Ratliff, Huttenlocher, & Cannon, 2012). Similarly, doll play, which is stereotypically associated with girls (Miller, Lurye, Zosuls, & Ruble, 2009), allows children to practice their socioemotional skills (Li & Wong, 2016). Because peers exclude children from activities based on gender stereotypes (Theimer et al., 2001), and because children become less interested in toys they think are "for" another gender (Martin, Eisenbud, & Rose, 1995), stereotyping may cause children to miss opportunities to develop important skills. Given its problematic nature, it is critical to develop methods to reduce stereotyping early in development.

Intervention approaches

One frequently tested approach for reducing gender stereotyping is counterstereotyping (Lenton, Bruder, & Sedikides, 2009; Pruden & Abad, 2013). Counterstereotyping interventions highlight specific examples of stereotype-inconsistent information. For example, a researcher hoping to alter

the belief that only women are nurses might expose participants to vignettes about male nurses. Although counterstereotyping is commonly used to combat adults' stereotypes (Lenton et al., 2009), there have been just a few informative interventions targeting children's gender stereotypes (e.g., for occupations, see Coyle & Liben, 2016; Scott & Feldman-Summers, 1979; Sherman & Zurbriggen, 2014; Steinke et al., 2007; Weeks & Porter, 1983; for musical instruments, see Pickering & Repacholi, 2001). Further, the results of such studies have been mixed (for a review, see Durkin, 1985): Some return positive effects (e.g., Pickering & Repacholi, 2001; Scott & Feldman-Summers, 1979; Sherman & Zurbriggen, 2014), and others do not (e.g., Coyle & Liben, 2016; Steinke et al., 2007; Weeks & Porter, 1983).

The failure of some counterstereotyping interventions aligns with the idea that it is generally easier to build up biases than to reduce them (Sherif, 1954). The challenge of reducing stereotyping is likely amplified when stereotypes accurately reflect the statistics in a child's environment (Jussim, Cain, Crawford, Harber, & Cohen, 2009; Swim, 1994). In the case of toy stereotypes, for example, boys *are* less likely than girls to play with dolls and girls *are* less likely than boys to play with vehicles (Servin, Bohlin, & Berlin, 1999).

Even among apparently successful interventions, mechanisms and sources of change can be difficult to discern. For instance, in a recent study focused on changing young children's toy stereotypes, children assigned to a counterstereotyping condition showed less stereotyping compared with those assigned to a stereotype-consistent condition (Spinner, Cameron, & Calogero, 2018). However, the study did not contain a stereotype-neutral condition or pretest; thus, it is difficult to determine whether the counterstereotyping intervention changed children's baseline thinking (for similar studies, see Ashton, 1983; Green, Bigler, & Catherwood, 2004; Pike & Jennings, 2005).

The current research

In the current research, we sought to improve on prior interventions focused on children's gender stereotyping by (a) choosing intervention strategies rooted in theories of stereotype acquisition and (b) carefully deconstructing our intervention approach over multiple conditions and studies in order to shed light on mechanisms underlying the observed intervention effects. Across four studies, we investigated whether counterstereotypical information changed children's gender stereotypes about toy preferences. We focused in particular on two highly stereotyped toys: dolls (associated with girls and not with boys) and trucks (associated with boys and not with girls) (Blakemore & Centers, 2005).

The current experiments targeted 5- and 6-year-old children (but also included 4- and 7-year-old children whose parents returned signed consent forms). This age range provides a particularly stringent test of the intervention's effectiveness because this is a time during development when children's gender stereotypes about toys are particularly rigid (Martin & Ruble, 2004, 2010; Weisgram, Fulcher, & Dinella, 2014). We did not have specific hypotheses based on participant age or gender; however, analyses focused on participant age and gender are available on OSF (https://osf.io/jdmqz/?view_only=a582d398d2264789aba318c9b1aee788).

We reasoned that to be successful, an intervention must be convincing enough to confront both existing messages in children's environments and children's own observations of what other children like. To this end, Experiment 1 participants received counterstereotypical messages with phrasing that has been theorized to encourage generalization of attributes within a category (i.e., language that bolsters stereotype acquisition); each counterstereotypical message contained gender category labels ("girls" and "boys"; see Arthur, Bigler, Liben, Gelman, & Ruble, 2008; Baron, Dunham, Banaji, & Carey, 2014; Bigler & Liben, 2007; Diesendruck & HaLevi, 2006; Patterson & Bigler, 2006; Rhodes & Gelman, 2008; Roberts, Ho, & Gelman, 2017; Waxman, 2010) and employed generic phrasing (e.g., "girls like trucks"; see Bian & Cimpian, 2017; Cimpian & Markman, 2008, 2011; Gelman, 2003; Rhodes, Leslie, & Tworek, 2012). An adult actor (described as a "teacher") delivered the intervention messages because children in our participants' age range trust information provided by adults and teachers (Corriveau & Harris, 2009; Jaswal, Croft, Setia, & Cole, 2010) and children commonly hear generic statements about gender categories from adults (Bigler & Liben, 2007; Endendijk et al., 2014; Gelman, Taylor, & Nguyen, 2004).

Experiment 1

Experiment 1 compared children's gender counterstereotypical beliefs about dolls and trucks after an intervention or control manipulation. Statements in both conditions contained gender group labels and generic phrasing. Participants in the intervention condition heard counterstereotypical statements about boys' and girls' toy preferences (e.g., "girls like trucks"), and participants in the control condition heard stereotype-irrelevant content (e.g., "girls like apples"). The control condition served to test whether simply seeing pictures of boys and girls and hearing sentences with generic phrasing and labels would affect children's beliefs about boys' and girls' liking of dolls and trucks.

Before and after they were exposed to the intervention or control manipulation, children answered questions about individuals and groups. *Individual measure* items asked children to rate individual children's liking of dolls or trucks. *Group measure* items asked children to rate groups of children's liking of dolls or trucks. We predicted that participants in the intervention condition would counterstereotype more at posttest than at pretest, and that the change in counterstereotyping would be larger for intervention condition participants than for control condition participants. We expected pretest-to-posttest changes to be greater when participants were answering *group measure* questions than when they were answering *individual measure* questions; even if the intervention did not alter participants' thinking about individual children's toy preferences, we reasoned that it may lead children to believe that a greater proportion of children like counterstereotypical toys.

Method

Participants

The participants were 49 children (27 girls; M_{age} = 5.80 years, range = 4.61–6.84). The planned sample size was 48 children, consistent with prior counterstereotyping intervention research with young children (e.g., Coyle & Liben, 2016; Weeks & Porter, 1983); however, one additional participant was scheduled, tested, and thus included in data analyses. Most participants were White (81.63%) and had at least one parent with a college degree (77.55%). In addition to the 49 participants included in analyses, 2 other children were tested but excluded from analyses because they made comments during the session that revealed their condition assignment to the experimenter (who was otherwise unaware of condition assignment). Children in Experiment 1 and all subsequent experiments were recruited and tested in the midwestern United States during 2016 and 2017 under a protocol approved by our institutional review board.

Measures and materials

Individual measure. In *individual measure* items, participants saw a line drawing of a doll (for boy target trials) or a truck (for girl target trials), followed by a photograph of a target child. They were asked to indicate (by pointing to one of three faces on a scale) how much the target child liked dolls or trucks (e.g., "How much does Remy like dolls?"). The scale ranged from "really likes" to "sort of likes" to "does not like" (see Fig. 1). We used uncommon gender-neutral names (as judged by research assistants in our laboratory) for target children because we believed they would be unfamiliar to most of our participants. Furthermore, using gender-neutral names allowed us to counterbalance across participants which names were paired with male versus female faces throughout the pretest and posttest.

Group measure. In *group measure* items, participants saw a line drawing of a doll (for boy target trials) or a truck (for girl target trials), followed by a scale containing eight faces (all girls or all boys) in three different arrangements in a horizontal orientation; on the left were eight faces with red "X" marks through each face, in the middle were the same eight faces with four faces circled in yellow, and on the right were the same eight faces all circled in yellow. Next, participants indicated (by pointing to a position on the scale) whether "none of them," "some of them," or "all of them" liked dolls (for boys) or trucks (for girls) (see Fig. 1).

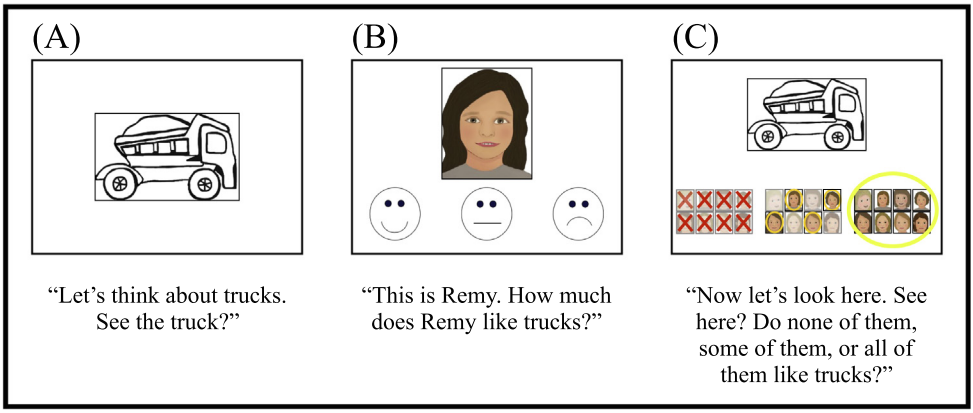

Fig. 1. Example displays from Experiment 1. (A) An example introduction to a target toy. (B) An example *individual measure* trial. (C) An example *group measure* trial. Because we did not have permission to publish the photographs of children who were used in the current experiments, this figure and all subsequent figures show an artist's rendition of the photographs.

Confidence measure. Following each answer using the *individual measure* and *group measure* scales, participants indicated how sure they were about their response. We probed participants' confidence in Experiments 1, 2, 3, and 4b to detect changes in participants' certainty in the event that our intervention did not work. However, as readers will see, most children changed their responses from pretest to posttest, rendering it difficult and unnecessary to analyze confidence measures. The confidence data and method are available on OSF (https://osf.io/jdmqz/?view_only=a582d398d2264789aba318c9b1aee788).

Teaching videos. Between pretest and posttest measures, participants viewed two teaching videos. Videos featured a woman (described as a "teacher") in front of a whiteboard. A woman played the teacher role because most preschool and elementary school teachers are female (Snyder, 2018). Participants saw one video focused on girls and trucks and another video focused on boys and dolls. In the intervention video focused on girls, the teacher said, "Look here—girls like trucks" four times while gesturing toward a group of two to four photographs of girls (totaling 12 photographs across the four trials). We included many photographs in order to match what participants saw to what the teacher was saying (i.e., the plural nouns "girls" and "boys"). We also varied the number of images presented in each of the four trials to maintain participants' interest in the displays. A black-and-white line drawing of a truck appeared underneath the photographs. A similar procedure occurred in the intervention teaching video focused on boys and dolls (see Fig. 2). Control condition teaching videos were identical to intervention condition videos except that the teacher said "girls like oranges [or apples]" and "boys like apples [or oranges]," with fruit pictures appearing underneath the photographs of children.

Procedure and design

A female experimenter tested all participants in a quiet private room at their school or in a university laboratory using a computer. Participants were randomly assigned to condition (intervention or control). Participants wore headphones during the teaching videos, and the computer screen was oriented so that the experimenter was unable to see the screen during the teaching videos; thus, she was unaware of condition assignment.

All experiment phases featured photographs of young children who were smiling and unfamiliar to participants. The pretest and posttest phases featured different photographs and names to evaluate whether the intervention influenced participants' inferences about new children of the same gender; no photographs or names were ever repeated during a session. We counterbalanced whether photographs appeared in the pretest or posttest across participants. All participants saw the same

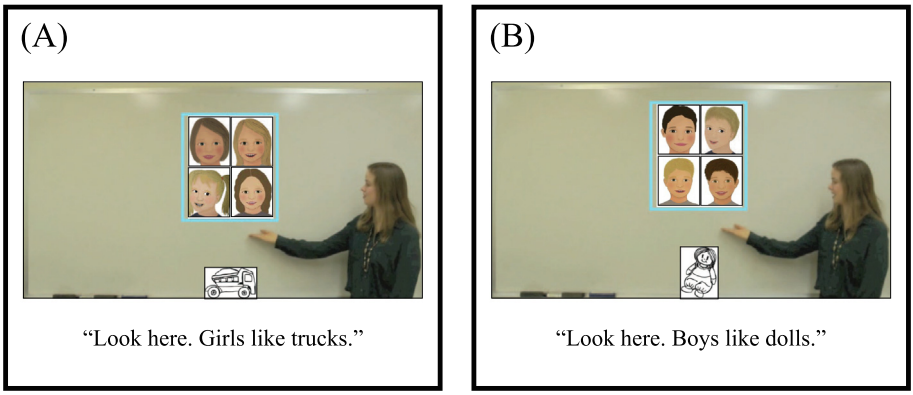

Fig. 2. Example intervention teaching videos. (A) An example snapshot from a girl–truck teaching video; (B) An example snapshot from a boy–doll teaching video.

photographs in the teaching videos, and these photographs were different from the photographs in the pretest and posttest phases.

Participants first learned how to use the scales for the measures. Next, participants completed two pretest blocks. One block focused on boys and dolls, and another block focused on girls and trucks (order counterbalanced across participants). Next, participants viewed a teaching video, followed by a first posttest. Finally, participants viewed a second teaching video, followed by a second posttest. The first teaching video and posttest focused on one gender (e.g., girls), and the second teaching video and posttest focused on the other gender (e.g., boys) (order counterbalanced across participants). Each pretest and posttest block consisted of four *individual measure* trials followed by one *group measure* trial. At the beginning of each posttest block, participants were told that they would be viewing new children they had never seen before to ensure that participants knew they were not being asked about children they had seen in a prior phase. See Fig. 3 for the procedural flow.

Scoring

"Counterstereotyping scores" were computed for each participant (one pretest set of scores and one posttest set of scores). For the *individual measure* trials, rating a target child as "really liking" the counterstereotypical toy was scored as 2, rating a child as "sort of liking" the counterstereotypical toy was scored as 1, and rating a child as "not liking" the counterstereotypical toy was scored as 0. Scores were then summed across the eight trials for each participant at pretest and posttest. Final counterstereotyping scores for the *individual measure* could range from 0 to 16 at pretest and from 0 to 16 at posttest. For the *group measure*, rating a target group by indicating "all of them" liked the counterstereotypical toy was scored as 2, rating a group by indicating that "some of them" liked the counterstereotypical toy was scored as 1, and rating a group by indicating that "none of them" liked the counterstereotypical toy was scored as 0. Scores were then summed across the two trials. Counterstereotyping scores for the *group measure* could range from 0 to 4 at pretest and from 0 to 4 at posttest.

Results

Analyses were conducted in R. Data files and R codes for all experiments are available on OSF (https://osf.io/jdmqz/?view_only=a582d398d2264789aba318c9b1aee788). Missing data for one cell was filled in using median imputation in R. Analyses in all studies evaluating condition differences as a change from baseline (pre/post) were conducted using an analysis of covariance (ANCOVA) with pretest scores serving as a covariate in the model. The ANCOVA approach for analyses of pre/post designs provides an unbiased estimate of effects and maximizes power in comparison with computing

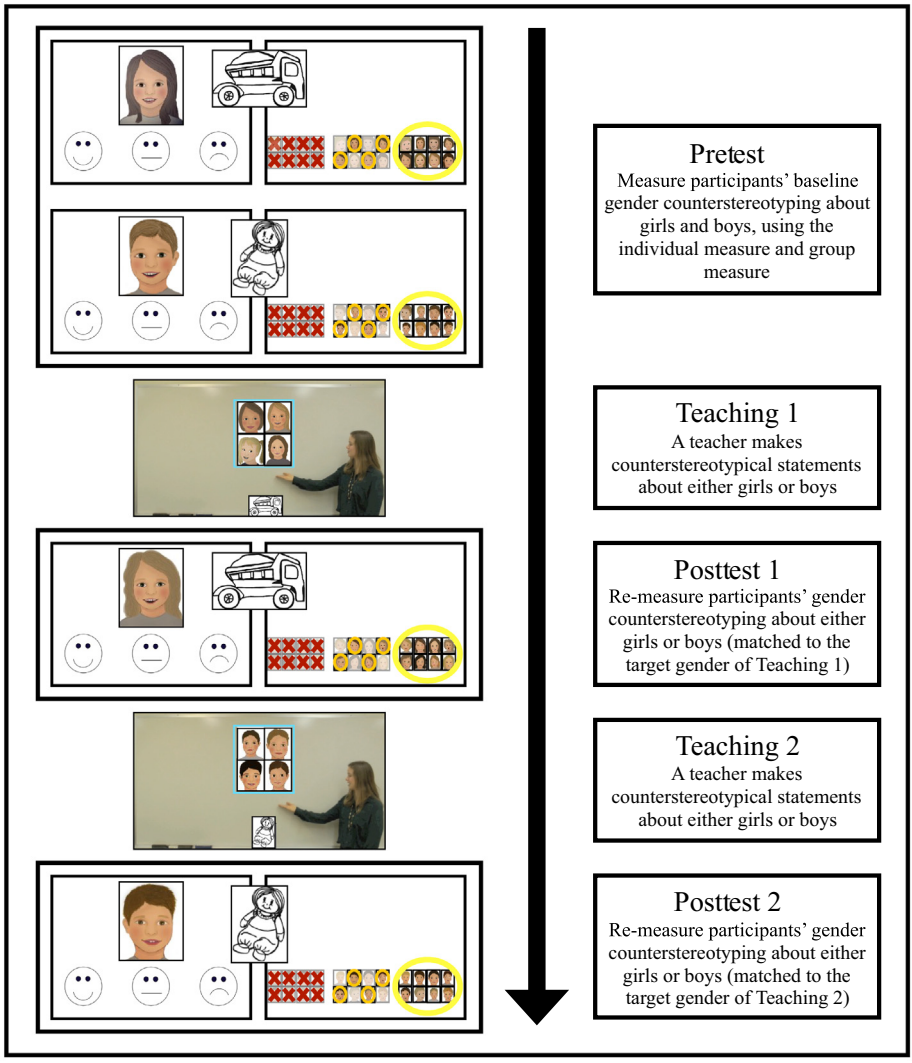

Fig. 3. Experiment 1 procedural flow. In this example, the first teaching video and posttest focused on girls and trucks, and the second teaching video and posttest focused on boys and dolls.

Table 1
Experiment 1 individual measure and group measure scores.

	Individual measure		Group measure	
Condition	Pretest	Posttest	Pretest	Posttest
Intervention	8.50 (3.83)	10.75 (4.55)	2.00 (1.02)	2.63 (1.24)
Control	6.84 (5.23)	5.32 (4.70)	1.48 (1.50)	1.76 (1.61)

Note. Means (and standard deviations) are shown. Higher scores indicate greater counterstereotyping.

a change score from pretest to posttest (see Van Breukelen, 2006, for a more detailed discussion of the analysis approach). See Table 1 for descriptive statistics.

Pretest

Regressing pretest counterstereotyping scores on condition showed that participants in the intervention and control conditions had similar pretest counterstereotyping scores on the *individual measure*, $F(1, 47) = 1.60$, $p = .21$, $\eta_p^2 = .03$, as well as on the *group measure*, $F(1, 47) = 1.99$, $p = .17$, $\eta_p^2 = .04$.

Posttest: Individual measure

Regressing posttest counterstereotyping scores on condition and pretest counterstereotyping scores showed that children's pretest counterstereotyping predicted their posttest counterstereotyping, regardless of condition (control = 0, intervention = 1), $F(1, 46) = 20.42$, $p < .001$, $\eta_p^2 = .31$. Critically, participants in the intervention condition showed a larger change in counterstereotyping from pretest to posttest than participants in the control condition, $F(1, 46) = 15.86$, $p < .001$, $\eta_p^2 = .26$. Furthermore, participants in the intervention condition had higher counterstereotyping scores at posttest than at pretest, $F(1, 23) = 5.56$, $p = .03$, $\eta_p^2 = .19$, whereas control participants had marginally *lower* counterstereotyping scores at posttest than at pretest, $F(1, 24) = 3.57$, $p = .07$, $\eta_p^2 = .13$.

Posttest: Group measure

Regressing posttest counterstereotyping scores on condition and pretest counterstereotyping scores revealed that pretest counterstereotyping scores significantly predicted posttest counterstereotyping scores, $F(1, 46) = 27.94$, $p < .001$, $\eta_p^2 = .38$, but there was no effect of condition, $F(1, 46) = 2.27$, $p = .14$, $\eta_p^2 = .05$.

Discussion

As predicted, participants in the intervention condition—but not participants in the control condition—increasingly expected girls to like trucks and boys to like dolls from pretest to posttest when responding to the *individual measure*. However, contrary to hypotheses, neither intervention nor control participants changed in their responding to the *group measure*. Young children have difficulty in reasoning about proportions (Boyer, Levine, & Huttenlocher, 2008) and in interpreting displays that connote proportions using discrete (rather than continuous) units (Jeong, Levine, & Huttenlocher, 2007). Therefore, children's performance on the *group measure* may reflect a misunderstanding of the particular scale used in this task. Alternatively, the contrasting results of the *group measure* and *individual measure* may reflect a promising change in children's thinking about toy stereotypes. The current intervention is unlikely to override children's existing knowledge that boys *are* statistically unlikely to like dolls (and girls to like trucks) at the group level (Servin et al., 1999), and this may explain the results of the *group measure*. However, the intervention may instead teach children that these group-level patterns do not necessarily reflect the preferences of individual boys and girls, as reflected by the results of the *individual measure*. Due to the ambiguity surrounding the interpretation of the null effects observed for the *group measure* in Experiment 1, in all following experiments we eliminated this measure and instead tested the boundary conditions of the effects observed for the *individual measure*.

The results of the *individual measure* in Experiment 1 provide evidence that the intervention changed children's thinking about toys and gender when considering the preferences of individual children. However, it is unclear whether the intervention's phrasing was critical to its success. Intervention messages contained noun labels, which highlight the presence and relevance of categories (e.g., Arthur et al., 2008; Prasada, 2000) and invite assumptions that individuals who share a label are alike (e.g., Rhodes & Gelman, 2008; Roberts et al., 2017; Waxman, 2010). Thus, the use of gender labels in the intervention may have been critical in promoting the generalization of information learned about one social category member to other category members (e.g., Baron et al., 2014; Diesendruck & HaLevi, 2006; Rhodes & Gelman, 2008; Roberts et al., 2017; Waxman, 2010).

In addition, the Experiment 1 intervention contained generic language. Generic statements express generalizations about a category (Cimpian & Markman, 2008; Gelman, 2003) and are particularly powerful because they express what is characteristic of a group rather than what is statistically common within a group, thereby remaining plausible in the face of exceptions (e.g., Bian & Cimpian, 2017; Gelman, 2003, 2004; Prasada, 2000). Generic counterstereotypical language may have been critical in

prompting participants to generalize the information provided in the intervention—even if the information contradicted their initial stereotypes and past experiences.

In Experiment 2, we manipulated the presence of generic language and gender labels in the intervention to disentangle each component's contribution to the effect observed in Experiment 1. Although much work has focused on the role of generic language and noun labels in creating or enhancing children's stereotypes (e.g., Bian & Cimpian, 2017; Bigler & Liben, 2007; Rhodes, 2012; Rhodes et al., 2012; Roberts et al., 2017; Waxman, 2010), little work has addressed the role of language in overturning children's stereotypes. Furthermore, no existing research we are aware of has compared whether counterstereotypical statements containing either (or both) of these linguistic features reduce children's stereotyping.

Experiment 2

Experiment 2 implemented a 2 (generic phrasing or demonstrative plural phrasing) × 2 (gender label or no gender label) between-participants design. As in Experiment 1, the experimenter was unaware of condition assignment and a teacher provided counterstereotypical information. We hypothesized that interventions featuring either generic language or gender labels would be more effective at increasing participants' counterstereotyping from pretest to posttest than an intervention containing neither because both linguistic features facilitate children's generalization within a social category. For the same reason, we also expected that the intervention containing both generic language and gender labels would have the strongest effect.

Method

Participants

Participants were 144 children (72 girls; M_{age} = 5.83 years, range = 4.70–7.12). Most participants were White (75.69%) and had at least one parent with a college degree (70.83%). In addition to the 144 participants included in data analyses, 3 other children were tested but excluded from data analyses because they did not complete the experiment (n = 2) or refused to wear headphones during the teaching videos (n = 1), which revealed the participant's condition assignment to the experimenter.

We conducted a power analysis in R using the effect size from Experiment 1 as our estimated effect size. This analysis revealed a suggested sample of 26 participants per condition. Given the changes in methodology from Experiment 1, we increased our sample size to 36 participants per condition in the event that we overestimated the effect size.

Materials, procedure, design, and scoring

The method was the same as in Experiment 1 except as follows. Participants were randomly assigned to one of four teaching conditions. The generic–gender label condition teaching video was identical to the Experiment 1 intervention video; the teacher said "boys like dolls" and "girls like trucks." In the demonstrative plural–gender label condition, the teacher said "these boys like dolls" and "these girls like trucks." In the generic–no gender label condition, the teacher said "kids like dolls" and "kids like trucks." In the demonstrative plural–no gender label condition, the teacher said "these kids like dolls" and "these kids like trucks."

Results

Missing data were imputed for four cells using median imputation. We regressed posttest counterstereotyping scores on generic condition (generic = 0.5, demonstrative plural = −0.5), gender label condition (gender label = 0.5, no gender label = −0.5), the interaction of generic and gender labels, and pretest counterstereotyping scores. Pretest counterstereotyping scores predicted posttest counterstereotyping scores, $F(1, 139) = 41.00$, $p < .001$, $\eta_p^2 = .23$. However, contrary to our hypotheses, posttest counterstereotyping scores were not influenced by generic language, $F(1, 139) = 1.05$, $p = .31$, $\eta_p^2 = .01$, gender labels, $F(1, 139) = 1.02$, $p = .32$, $\eta_p^2 = .01$), or the interaction of generic language and

Table 2
Individual measure scores.

Experiment	Condition	Pretest	Posttest
1	Intervention	8.50 (3.83)	10.75 (4.55)
	Control	6.84 (5.23)	5.32 (4.70)
2	Generic–gender label	8.31 (3.93)	11.61 (4.30)
	Generic–no gender label	6.17 (4.90)	9.31 (5.29)
	Demonstrative plural–gender label	7.86 (5.35)	11.47 (3.92)
	Demonstrative plural–no gender label	6.78 (4.42)	10.92 (4.27)
3	Demonstrative singular–no gender label	8.22 (5.68)	11.44 (5.23)
4b	Generic–gender label	8.06 (4.05)	11.97 (3.79)
	Single exemplar	7.47 (5.12)	9.14 (4.77)
	Four exemplars	6.44 (4.93)	11.19 (4.59)

Note. Means (and standard deviations) are shown. Higher scores indicate greater counterstereotyping.

gender labels, $F(1, 139) = 0.86$, $p = .36$, $\eta_p^2 = .01$. In fact, regressing the difference between posttest and pretest counterstereotyping scores in each condition on 1 revealed that participants' scores increased from pretest to posttest in all four conditions: demonstrative plural–gender label, $F(1, 35) = 21.77$, $p < .001$, $\eta_p^2 = .38$; demonstrative plural–no gender label, $F(1, 35) = 22.48$, $p < .001$, $\eta_p^2 = .39$; generic–gender label, $F(1, 35) = 31.90$, $p < .001$, $\eta_p^2 = .48$; and generic–no gender label, $F(1, 35) = 14.07$, $p < .001$, $\eta_p^2 = .29$. (Note that regressing difference scores on 1 is equivalent to conducting a paired-samples *t*-test evaluating the difference between pretest and posttest. A significant intercept in the model indicates a significant difference between pretest and posttest scores.) See Table 2 for means and standard deviations.

Discussion

Experiment 2 provides further evidence that a brief intervention can change children's toy stereotyping. Contrary to hypotheses, children in all conditions were more likely to report that girls like trucks and boys like dolls at posttest than at pretest, and there were no condition differences. What might explain the lack of condition differences? One possibility is that the language used to deliver counterstereotypical messages was plural in every condition (e.g., "girls," "these kids"). Learning about groups via language that connotes multiple individuals might help children to learn and generalize counterstereotypical information more readily than we predicted. Alternatively, children might not need to hear counterstereotypical information via language connoting groups (i.e., generic phrasing, group labels, or plural phrasing) in order to learn and apply intervention messages; individual exemplars may be sufficient. Experiment 3 aimed to distinguish these possibilities.

Experiment 3

Experiment 3 employed an intervention that communicated counterstereotypical information via demonstrative singular phrasing without gender group labels (e.g., "this kid likes trucks"). The experimenter was unaware of condition assignment, the content and number of conditions (one condition), and hypotheses.

Method

Participants

We tested 36 children (16 girls; M_{age} = 5.74 years, range = 5.19–6.47). Most participants were White (69.44%) and had at least one parent with a college degree (55.56%). In addition to the 36 participants included in data analyses, 1 additional child was tested but excluded from analyses due to experimenter error.

Materials, procedure, design, and scoring

The method was identical to that in Experiment 2 except that all participants were assigned to one condition (demonstrative singular–no gender label). Participants heard "this kid likes trucks" four times when learning about girls and "this kid likes dolls" four times when learning about boys. Participants viewed the same visual displays from Experiments 1 and 2, featuring four groups of children per video ranging from 2 to 4 children per group. Because the teacher gestured toward the images (as she had done in Experiments 1 and 2), it appeared feasible that she was gesturing toward just one individual within each group.

Results

Regressing the difference between posttest and pretest counterstereotyping scores on 1 revealed that participants had higher individual counterstereotyping scores at posttest than at pretest, $F(1, 35) = 12.58$, $p = .001$, $\eta_p^2 = .26$. See Table 2. We also analyzed the data from Experiments 2 and 3 together to detect condition differences. We created four dummy codes for the conditions using the condition in Experiment 3 (demonstrative singular–no gender label) as the reference group. Regressing posttest counterstereotyping scores on the dummy codes and pretest counterstereotyping scores showed that pretest counterstereotyping scores predicted posttest counterstereotyping scores, $F(1, 173) = 51.95$, $p < .001$, $\eta_p^2 = .23$. However, there were no condition effects (all ps > .250).

Discussion

The intervention tested in Experiment 3 was successful, and we observed no differences between conditions in Experiments 2 and 3. These findings raise two questions. First, can this brief laboratory intervention change children's *behavior,* or are its effects limited to just performance on the *individual measure*—a measure of children's self-reported beliefs? Second, is it possible that the number of individual children featured in the teaching videos of Experiments 1 through 3—12 children in total—promoted children's learning and application of counterstereotypical information?

Experiment 4 addressed these questions. First, in Experiment 4a, we created and tested a novel behavioral *toy selection measure* in which participants select a toy for another child to receive. Although prior research using similar tasks has shown that children choose stereotypical toys at baseline (e.g., Eisenberg et al., 1982), we first sought to confirm that our measure would replicate this well-established pattern (Experiment 4a). Next, in Experiment 4b, we used this novel measure along with the *individual measure* to continue testing the current intervention's effectiveness with fewer counterstereotypical exemplars per teaching video.

Experiment 4a

Method

Participants

We tested 36 children (18 girls; M_{age} = 5.45 years, range = 4.05–6.85). Most children were White (72.22%) and had at least one parent with a college degree (58.33%).

Materials, procedure, and design

Participants saw a photograph of an unfamiliar target child in front of three toys: a baby doll in yellow clothing, a yellow and red dump truck, and a yellow maraca. Half of the male (and half of the female) participants selected a toy for a female target, and half of the male (and half of the female) participants selected a toy for a male target.

A White female experimenter tested participants in a quiet space at their school or in a university laboratory. Participants were told that the target child was visiting later and would be given only one toy to play with. Participants were then told to select which toy the target child would receive by placing the toy they believed the target child would like most in a black bin. The experimenter did not

label the toys or the target child's gender; rather, the experimenter placed the target child's photograph and the toy options on a table, provided the instructions, and then averted her gaze while participants made their selection.

For exploratory purposes, after participants placed their first selection in the bin, the experimenter asked children to select a second toy for the target. We focus only on participants' first toy selections in our analyses and discussion here (see OSF [https://osf.io/jdmqz/?view_only=a582d398d2264789aba318c9b1aee788] for ranked selections).

Results and discussion

We conducted a chi-square analysis to determine whether the distribution of participants' choices was significantly different from chance (one third choosing each toy). For this analysis, counterstereotypical choices (selecting a doll for a boy or a truck for a girl) were scored as −1, stereotypical choices (selecting a doll for a girl or a truck for a boy) were scored as 1, and selections of the maraca were scored as 0. As expected, participants chose more stereotypical toys than would be expected by chance ($n_{\text{stereotypical}} = 28$, $n_{\text{counterstereotypical}} = 5$, $n_{\text{maraca}} = 3$), $\chi^2(2) = 32.17$, $p < .001$, most often choosing dolls for girls and trucks for boys. In Experiment 4b, we used this measure, in addition to the *individual measure* from Experiments 1 to 3, to test the impact of three intervention conditions.

Experiment 4b

Participants in Experiment 4b were randomly assigned to one of three intervention conditions. One condition used the generic–gender label teaching videos from Experiments 1 and 2. The other conditions featured either four (four exemplars condition) or one (single exemplar condition) individual counterstereotypical exemplars per teaching video. The experimenter was unaware of condition assignment.

We predicted that participants in all conditions would counterstereotype more at posttest than at pretest. However, we expected that participants in the generic–gender label condition would show the greatest change in counterstereotyping from pretest to posttest, followed by the four exemplars and single exemplar conditions, respectively. We expected that the combination of generic language, group labels, and displays featuring groups of children in the generic–gender label condition would support participants' learning of counterstereotypes. On the other hand, we expected that the emphasis on individuals in the single exemplar and four exemplars conditions would undermine children's learning at the group level, thereby resulting in less group-level learning as the number of exemplars decreased.

Method

Participants

Participants were 108 children (54 girls; $M_{\text{age}} = 5.54$ years, range = 4.09–7.07) randomly assigned to one of three conditions. In addition to the 108 participants included in data analyses, another 4 children were tested but excluded from analyses because they did not complete the experiment ($n = 3$) or made comments that made the experimenter aware of condition assignment ($n = 1$). Most participants were White (76.85%) and had at least one parent with a college degree (73.15%).

Materials, procedure, design, and scoring

The method was the same as in Experiments 2 and 3 (for pretest, teaching videos, and posttest) and Experiment 4a (for the *toy selection measure*) except for the following changes: Participants were assigned to one of three conditions. One condition featured the teaching videos from the generic–gender label condition of Experiments 1 and 2, and the remaining two conditions featured novel teaching videos. In the single exemplar condition, one child was featured in each teaching video; participants heard "Harumi likes trucks" four times while viewing an image of the same girl each time (for the teaching video focused on girls) and "Izumi likes dolls" four times while viewing an image of the same boy each time (for the teaching video focused on boys). In the four exemplars condition, four different children were featured in each teaching video; participants heard "Yuki likes trucks" with an image of

one girl, "Yoval likes trucks" with an image of another girl, "Tomomi likes trucks" with an image of another girl, and "Harumi likes trucks" with an image of another girl (for the teaching video focused on girls). The same procedure was used for the teaching video focused on boys with different names (Izumi, Riku, Yahli, and Misumi) and images of boys. No names or images were repeated in different phases of the experiment.

After completing both *individual measure* posttest phases, participants re-viewed the first teaching video, which focused either on girl(s) and trucks or on boy(s) and dolls (reminder trial). Next, participants completed the *toy selection measure*, identical to that in Experiment 4a. The target in the *toy selection measure* was gender-matched to the target(s) featured in the first teaching video that participants viewed (and, thus, was gender-matched to the target(s) featured in the reminder trial). Half of male participants chose a toy for a male target, and half of male participants chose a toy for a female target; the same was true of female participants.

Results

Individual measure

Regressing posttest counterstereotyping scores on a contrast of interest (generic–gender label = 1, four exemplars = 0, single exemplar = −1), a contrast testing the residual between-group variance (−1, 2, −1), and pretest counterstereotyping scores showed that, regardless of condition, participants' pretest counterstereotyping scores predicted their posttest counterstereotyping scores, $F(1, 104) = 19.49$, $p < .001$, $\eta_p^2 = .16$. The contrast of interest (1, 0, −1) was statistically significant, $F(1, 104) = 7.46$, $p < .01$, $\eta_p^2 = .07$, whereas the contrast testing the residual between-group variance (−1, 2, −1) was not, $F(1, 104) = 1.82$, $p = .18$, $\eta_p^2 = .02$. Participants in the generic–gender label condition had higher counterstereotyping scores ($M = 11.97$, $SD = 3.79$) at posttest than participants in the four exemplars condition ($M = 11.19$, $SD = 4.59$), who had higher counterstereotyping scores at posttest than participants in the single exemplar condition ($M = 9.14$, $SD = 4.77$). Regressing the difference between posttest and pretest counterstereotyping scores in each condition on 1 showed that counterstereotyping scores increased from pretest to posttest in all three conditions: generic–gender label, $F(1, 35) = 17.92$, $p < .001$, $\eta_p^2 = .34$; four exemplars, $F(1, 35) = 32.42$, $p < .001$, $\eta_p^2 = .48$; and single exemplar, $F(1, 35) = 5.09$, $p = .03$, $\eta_p^2 = .13$. See Table 2 for means and standard deviations.

Toy selection measure

Participants in the single exemplar and four exemplars conditions selected stereotypical toys at rates similar to participants in Experiment 4a (single exemplar: stereotypical = 25, counterstereotypical = 8, maraca = 3; four exemplars: stereotypical = 24, counterstereotypical = 7, maraca = 5). Neither of these distributions of scores differs from the distribution of scores observed in Experiment 4a: single exemplar, $\chi^2(2) = 2.12$, $p = .35$; four exemplars, $\chi^2(2) = 2.70$, $p = .26$. However, participants in the generic–gender label condition selected counterstereotypical toys at higher rates, and stereotypical toys at lower rates, than participants in Experiment 4a and participants in either of the other two conditions of Experiment 4b ($n_{stereotypical} = 20$, $n_{counterstereotypical} = 14$, $n_{maraca} = 2$). This distribution of scores differed from the distribution of scores observed in Experiment 4a, $\chi^2(2) = 18.82$, $p < .001$.

We conducted a multinomial logistic regression to compare participants' scores in each condition of Experiment 4b with the scores observed in Experiment 4a. Stereotypical toy selections were coded as the reference group. Children in the generic–gender label condition—but not children in other conditions—were approximately four times more likely to select a counterstereotypical toy for a peer than children who did not participate in an intervention (Experiment 4a, reference group), odds ratio for counterstereotypical toy selection = 3.92, odds ratio for maraca toy selection = 0.93, $SE = 0.60$, $p = .02$; other conditions, $ps > .35$.

Discussion

Children in the generic–gender label condition demonstrated robust change on the *individual measure* from pretest to posttest, and they were also more likely to make counterstereotypical toy

selections for peers than children in the baseline condition (Experiment 4a). Participants in the single exemplar and four exemplars conditions also changed their responses about other children's toy preferences on the *individual measure* but did so less robustly than children in the generic–gender label condition; furthermore, children's toy selections in these conditions were indistinguishable from those of Experiment 4a participants. We discuss these findings further below.

General discussion

Summary of effects

The current research provides consistent support for the efficacy of counterstereotyping as a means to change children's thinking about boys' and girls' toy preferences. From pretest to posttest in every intervention condition (see Table 3 for a summary), children became more likely to indicate that boys liked dolls and that girls liked trucks. The only condition where this change was not observed was the control condition of Experiment 1 in which participants learned stereotype-irrelevant content.

The primary goal of the current research was to develop a strategy for reducing children's stereotypical beliefs about peers' toy preferences. A secondary goal was to test the intervention's limits and shed light on mechanisms of change by making incremental changes to the language and visual displays used in the design. Experiment 4b is the only case where some intervention versions were more effective than others: Pretest-to-posttest change on the *individual measure* was statistically significant in all three conditions, but change was smaller in the single exemplar condition than in the four exemplars condition and was most robust in the generic–gender label condition. Condition differences in Experiment 4b were also evident on the *toy selection measure*. In particular, only participants in the generic–gender label condition were more likely to select counterstereotypical toys for peers than participants who had not viewed an intervention (Experiment 4a), and children's behavior in the remaining two conditions did not differ from that of Experiment 4a participants.

Interpreting similarities and differences across conditions

The language used to describe a social category and its members affects children's learning about the category (e.g., Roberts et al., 2017). For example, saying "Hibbles eat berries" leads children to generalize berry-eating throughout the category of Hibbles, whereas saying "this one eats berries" while showing an image of a Hibble does not (Roberts et al., 2017). Indeed, this is precisely why we employed generic language and gender labels in Experiment 1: The presence of these factors enhances

Table 3
Features of teaching videos by experiment and condition.

Experiment	Condition	Teacher statements (M)	Teacher statements (F)	Photos
1	Intervention	"Boys like dolls"	"Girls like trucks"	12
	Control	"Boys like apples"	"Girls like oranges"	12
2	Generic–gender label	"Boys like dolls"	"Girls like trucks"	12
	Generic–no gender label	"Kids like dolls"	"Kids like trucks"	12
	Demonstrative plural–gender label	"These boys like dolls"	"These girls like trucks"	12
	Demonstrative plural–no gender label	"These kids like dolls"	"These kids like trucks"	12
3	Demonstrative singular–no gender label	"This kid likes dolls"	"This kid likes trucks"	12
4b	Generic–gender label	"Boys like dolls"	"Girls like trucks"	12
	Four exemplars	"Izumi likes dolls, Riku likes dolls. . ."	"Yuki likes trucks, Yoval likes trucks. . ."	4
	Single exemplar	"Izumi likes dolls"	"Harumi likes trucks"	1

Note. Total numbers of photos of each gender used in teaching phase videos are shown. Participants viewed two teaching videos (one about boys and one about girls).

stereotyping in children within a single experimental session (see Rhodes et al., 2012; Roberts et al., 2017; Waxman, 2010), so we reasoned that these factors may also be necessary to teach children counterstereotypical information over the course of a short intervention. However, all intervention conditions of Experiments 2 and 3—where children learned, for example, that "boys like dolls" or "this kid likes dolls"—were equally effective. These results diverge from the stereotype acquisition literature in which individual exemplars are not typically sufficient to learn new stereotypes (Rhodes et al., 2012; Roberts et al., 2017). What explains the intervention's consistent effectiveness across conditions in Experiments 2 and 3?

One possibility is that common features across all intervention conditions led participants to rapidly pick up on the presented patterns (i.e., linking boys with dolls and girls with trucks). Although the language used to convey these patterns differed between conditions, the patterns themselves did not differ. Thus, factors that were consistent across conditions, such as message repetition (four times per video), may have played an important role in bolstering children's learning.

A second possibility is that children in every condition remembered the intervention messages similarly. By 4 years of age, children automatically encode people's genders even when gender is not mentioned; after a time delay, they tend to remember a person's gender but not necessarily individuating information such as facial features (Weisman, Johnson, & Shutts, 2015). However, there is no evidence that children automatically encode novel social categories, which are commonly used in stereotype acquisition literature, such as Hibbles and Glerks (Roberts et al., 2017). Thus, the fact that we intervened on existing categories to which children are automatically attuned may explain the divergence between our intervention and the stereotype acquisition literature; even when participants *heard* "this kid likes dolls," they may have *encoded* the message as "boys like dolls," leading them to extend the intervention message ("like dolls") to all members of the category ("boys").

In Experiment 4b, changes from pretest to posttest were most robust in the generic–gender label condition, followed by the four exemplars and single exemplar conditions, respectively. What in particular about the single exemplar and four exemplars interventions undermined their success? One possibility is that children in these conditions subtyped intervention exemplars. Subtyping occurs when people encounter counterstereotypical exemplars and subcategorize these individuals as unrepresentative exceptions rather than revising their stereotypes (see Hewstone, 1994, for a review). As the number of "exceptions" increases, people eventually begin incorporating the counterstereotypical information into their understanding of the group (Johnston & Hewstone, 1992). If participants in the single exemplar and four exemplars conditions engaged in subtyping, then this could explain why children who saw four exemplars per teaching video changed more from pretest to posttest than those who saw one exemplar per video.

Beyond the effects observed on the *individual measure* of Experiment 4b, there were also condition differences in the toy selection task: Only generic–gender label condition participants made more counterstereotypical choices (and fewer stereotypical choices) than participants who had not participated in an intervention. These results are particularly striking because participants never learned that boys and girls dislike stereotypical toys (e.g., that girls dislike dolls), yet many of them selected counterstereotypical toys *over* stereotypical ones. Why did the *toy selection measure* reveal starker condition differences than the *individual measure*? The toy selection task may be a stricter test because it captures whether children's self-reported attitude changes are reflected in their behaviors. Alternatively, given the general asymmetry between people's attitudes and behaviors (Ajzen & Fishbein, 1977), the *toy selection measure* may instead capture qualitatively different effects than the *individual measure*.

Mechanisms of effects and future directions

What mechanisms underlie the changes we observed in participants' responses? It is possible that the intervention weakened participants' linking of gender and toy preferences because each intervention provided children with evidence that girls (or a girl) like(s) trucks and that boys (or a boy) like(s) dolls. Thus, at posttest, participants may have applied the latest information they had on the topic—that girls like trucks and boys like dolls. Another possibility—not incompatible with the first—is that the intervention undermined the utility of gender as a cue to how much children like *any* objects.

Future studies could investigate this question by determining whether the effects extend to children's beliefs about other stereotyped toys (e.g., tea set), abilities (e.g., math), or occupations (e.g., nurse).

Another question concerns whether participants altered their responses in an attempt to please the experimenter but did not actually come to hold different beliefs about boys' and girls' toy preferences. One point that casts doubt on this proposal is that younger and older children in the current research performed similarly (see OSF [https://osf.io/jdmqz/?view_only=a582d398d2264789aba318c9b1aee788] for analyses). Although the age at which children begin to adjust their behaviors to please others varies by context, the motivation to behave in accordance with adults' expectations grows as children age (e.g., Fitzroy & Rutland, 2010; Fu & Lee, 2007; Shaw et al., 2014; Silver & Shaw, 2018). Yet, younger children in our sample were as likely as older children to show effects as a result of the intervention. Furthermore, children's behaviors differed by condition on the *toy selection measure*—a task that requires more than simply responding to an experimenter's questions about boys' and girls' toy preferences.

Even if children in the current research changed their behavior according to what they thought the experimenter wanted them to do, the resulting behavior is still relevant from an intervention standpoint. In a classroom or play context, similar behavioral changes could mean that boys and girls are included in activities from which they would have typically been excluded. It is also worth noting that changes in short-term behavior can affect attitudes and beliefs through dissonance-reducing processes (Bem, 1972; Festinger & Carlsmith, 1959). When children's behaviors (e.g., giving trucks to girls) diverge from their beliefs (e.g., believing girls dislike trucks), this conflict can be resolved if children adjust their beliefs to align with their behaviors. As one example of this process in another domain, children initially follow novel norms based on social motivations (e.g., approval from others) but eventually internalize the norms (e.g., they feel guilty if they violate them; Jensen, Vaish, & Schmidt, 2014). If participants in the current research aligned their responses with perceived injunctive norms expressed by the experimenter, then with continued exposure to the intervention they may incorporate these norms into their personal beliefs. Future research could test this possibility using a longitudinal approach.

In future efforts, it will be imperative to determine the longevity and generalizability of the current intervention's effects. For example, future research should establish whether the current method, which focused on dolls and trucks, would affect children's beliefs about toys not discussed during the intervention (e.g., stereotypes about toy ponies and toy guns) or children's beliefs in non-toy domains (e.g., stereotypes about occupations). In addition, the participant samples in the current experiments and in the studies from which we drew our theoretical framework were fairly homogeneous in terms of race, socioeconomic status (SES), and geography. Future research should determine whether the effects observed here extend to children from cultural contexts outside of primarily White, middle- to high-SES children in the United States. Furthermore, although the results of Experiment 4's *toy selection measure* provide preliminary evidence that the intervention's effects extend beyond the context of the intervention, it will be important to test this possibility in more ecologically valid contexts. Finally, participants in the current experiments completed the posttest measure immediately after viewing the intervention videos; thus, it would also be informative in future research using this intervention strategy to delay the administration of the posttest after the intervention in order to evaluate the duration of effects and to shed light on condition differences that might emerge only after a time delay.

In the current research, we taught children counterstereotypical information directly, reasoning that such a strategy would be necessary and effective for our participant sample—namely, children reared in environments where gender categories are emphasized and children who already use gender categories to make predictions about others. It is worth noting, however, that other approaches may be more effective or appropriate for children of different ages or in different cultural contexts. For example, among children who do not already possess gender stereotypes, preventative approaches (e.g., early exposure to counterstereotypes in media) may be more effective. In cases where children are reared in environments that deemphasize gender (e.g., gender-neutral classrooms; Shutts, Kenward, Falk, Ivegran, & Fawcett, 2017), addressing gender stereotypes by highlighting and commenting on other people's gender would not be appropriate. In future research, it will be important to determine when and for whom particular intervention strategies are most effective.

Acknowledgments

This research was supported by a Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grant (R01 HD070890) to K.S. and National Science Foundation (NSF) Graduate Research Fellowships to R.A.K. and M.P.R. This study was also supported in part by a core grant to the Waisman Center from the NICHD (U54 HD090256). We thank Bailey Immel for assistance with data collection. Thanks also to Patricia G. Devine, Katherine D. Kinzler, and Tory Ash for helpful comments on prior versions of the manuscript. Portions of this research were presented at the 2017 Society for Research in Child Development biennial meeting, Austin, TX.

References

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*, 888–918.
Arnold, L., Seidl, M., & Deloney, A. (2015). Hegemony, gender stereotypes and Disney: A content analysis of Frozen and Snow White. *Concordia Journal of Communication Research, 2*, 1–24.
Arthur, A. E., Bigler, R. S., Liben, L. S., Gelman, S. A., & Ruble, D. N. (2008). Gender stereotyping and prejudice in young children. In S. R. Levy & M. Killen (Eds.), *Intergroup attitudes and relations in childhood through adulthood* (pp. 66–86). New York: Oxford University Press.
Ashton, E. (1983). Measures of play behavior: The influence of sex-role stereotyped children's books. *Sex Roles, 9*, 43–47.
Auster, C. J., & Mansbach, C. S. (2012). The gender marketing of toys: An analysis of color and type of toy on the Disney store website. *Sex Roles, 67*, 375–388.
Baron, A. S., Dunham, Y., Banaji, M., & Carey, S. (2014). Constraints on the acquisition of social category concepts. *Journal of Cognition and Development, 15*, 238–268.
Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 1–62). New York: Academic Press.
Berry, T., & Wilkins, J. (2017). The gendered portrayal of inanimate characters in children's books. *Journal of Children's Literature, 43*, 4–15.
Bian, L., & Cimpian, A. (2017). Are stereotypes accurate? A perspective from the cognitive science of concepts. *Behavioral and Brain Sciences, 40*, 22–24.
Bigler, R. S., & Liben, L. S. (2007). Developmental intergroup theory: Explaining and reducing children's social stereotyping and prejudice. *Current Directions in Psychological Science, 16*, 162–166.
Blakemore, J. E. O., & Centers, R. E. (2005). Characteristics of boys' and girls' toys. *Sex Roles, 53*, 619–633.
Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology, 44*, 1478–1490.
Cherney, I. D., & Dempsey, J. (2010). Young children's classification, stereotyping and play behaviour for gender neutral and ambiguous toys. *Educational Psychology, 30*, 651–669.
Cimpian, A., & Markman, E. M. (2008). Preschool children's use of cues to generic meaning. *Cognition, 107*, 19–53.
Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development, 82*, 471–492.
Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science, 12*, 426–437.
Cowan, G., & Hoffman, C. D. (1986). Gender stereotyping in young children: Evidence to support a concept-learning approach. *Sex Roles, 14*, 211–224.
Coyle, E. F., & Liben, L. S. (2016). Affecting girls' activity and job interests through play: The moderating roles of personal gender salience and game characteristics. *Child Development, 87*, 414–428.
Davis, S. N. (2003). Sex stereotypes in commercials targeted toward children: A content analysis. *Sociological Spectrum, 23*, 407–424.
Diesendruck, G., & HaLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. *Child Development, 77*, 539–553.
Durkin, K. (1985). Television and sex-role acquisition: 3. Counter-stereotyping. *British Journal of Social Psychology, 24*, 211–222.
Eisenberg, N., Murray, E., & Hite, T. (1982). Children's reasoning regarding sex-typed toy choices. *Child Development, 53*, 81–86.
Endendijk, J. J., Groeneveld, M. G., van der Pol, L. D., van Berkel, S. R., Hallers-Haalboom, E. T., Mesman, J., & Bakermans-Kranenburg, M. J. (2014). Boys don't play with dolls: Mothers' and fathers' gender talk during picture book reading. *Parenting, 14*, 141–161.
Etaugh, C., & Liss, M. B. (1992). Home, school, and playroom: Training grounds for adult gender roles. *Sex Roles, 26*, 129–147.
Fabes, R. A., Martin, C. L., & Hanish, L. D. (2003). Young children's play qualities in same-, other-, and mixed-sex peer groups. *Child Development, 74*, 921–932.
Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203–210.
Fitzroy, S., & Rutland, A. (2010). Learning to control ethnic intergroup bias in childhood. *European Journal of Social Psychology, 40*, 679–693.
Freeman, N. K. (2007). Preschoolers' perceptions of gender appropriate toys and their parents' beliefs about genderized behaviors: Miscommunication, mixed messages, or hidden truths? *Early Childhood Education Journal, 34*, 357–366.
Fu, G., & Lee, K. (2007). Social grooming in the kindergarten: The emergence of flattery behavior. *Developmental Science, 10*, 255–265.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought (Oxford Series in Cognitive Development)*. New York: Oxford University Press.
Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences, 8*, 404–409.
Gelman, S. A., Taylor, M. G., & Nguyen, S. P. (2004). Mother–child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development, 69*(1), 1–142.
Gooden, A. M., & Gooden, M. A. (2001). Gender representation in notable children's picture books: 1995–1999. *Sex Roles, 45*, 89–101.
Green, V. A., Bigler, R., & Catherwood, D. (2004). The variability and flexibility of gender-typed toy play: A close look at children's behavioral responses to counterstereotypic models. *Sex Roles, 51*, 371–386.
Halim, M. L., Ruble, D., Tamis-LeMonda, C., & Shrout, P. E. (2013). Rigidity in gender-typed behaviors in early childhood: A longitudinal study of ethnic minority children. *Child Development, 84*, 1269–1284.
Halpern, H. P., & Perry-Jenkins, M. (2016). Parents' gender ideology and gendered behavior as predictors of children's gender-role attitudes: A longitudinal exploration. *Sex Roles, 74*, 527–542.
Hewstone, M. (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology, 5*, 69–109.
Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science, 21*, 1541–1547.
Jensen, K., Vaish, A., & Schmidt, M. F. (2014). The emergence of human prosociality: Aligning with others through feelings, concerns, and norms. *Frontiers in Psychology, 5*, 822–837.
Jeong, Y., Levine, S. C., & Huttenlocher, J. (2007). The development of proportional reasoning: Effect of continuous versus discrete quantities. *Journal of Cognition and Development, 8*, 237–256.
Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: III. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology, 28*, 360–386.
Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 99–227). New York: Psychology Press.
Kahlenberg, S. G., & Hein, M. M. (2010). Progression on Nickelodeon? Gender-role stereotypes in toy commercials. *Sex Roles, 62*, 830–847.
Kane, E. W. (2006). "No way my boys are going to be like that!" Parents' responses to children's gender nonconformity. *Gender & Society, 20*, 149–176.
Leaper, C., Breed, L., Hoffman, L., & Perlman, C. A. (2002). Variations in the gender-stereotyped content of children's television cartoons across genres. *Journal of Applied Social Psychology, 32*, 1653–1662.
Lenton, A. P., Bruder, M., & Sedikides, C. (2009). A meta-analysis on the malleability of automatic gender stereotypes. *Psychology of Women Quarterly, 33*, 183–196.
Levine, S. C., Ratliff, K. R., Huttenlocher, J., & Cannon, J. (2012). Early puzzle play: A predictor of preschoolers' spatial transformation skill. *Developmental Psychology, 48*, 530–542.
Li, R. Y. H., & Wong, W. I. (2016). Gender-typed play and social abilities in boys and girls: Are they related?. *Sex Roles, 74*, 399–410.
MacPhee, D., & Prendergast, S. (2019). Room for improvement: Girls' and boys' home environments are still gendered. *Sex Roles, 80*, 332–346.
Martin, C. L., Eisenbud, L., & Rose, H. (1995). Children's gender-based reasoning about toys. *Child Development, 66*, 1453–1471.
Martin, C. L., & Ruble, D. (2004). Children's search for gender cues: Cognitive perspectives on gender development. *Current Directions in Psychological Science, 13*, 67–70.
Martin, C. L., & Ruble, D. N. (2010). Patterns of gender development. *Annual Review of Psychology, 61*, 353–381.
Miller, C. L. (1987). Qualitative differences among gender-stereotyped toys: Implications for cognitive and social development in girls and boys. *Sex Roles, 16*, 473–487.
Miller, C., Lurye, L. E., Zosuls, K. M., & Ruble, D. N. (2009). Accessibility of gender stereotype domains: Developmental and gender differences in children. *Sex Roles, 60*, 870–881.
Murnen, S. K., Greenfield, C., Younger, A., & Boyd, H. (2016). Boys act and girls appear: A content analysis of gender stereotypes associated with characters in children's popular culture. *Sex Roles, 74*, 78–91.
Patterson, M. M., & Bigler, R. S. (2006). Preschool children's attention to environmental messages about groups: Social categorization and the origins of intergroup bias. *Child Development, 77*, 847–860.
Pickering, S., & Repacholi, B. (2001). Modifying children's gender-typed musical instrument preferences: The effects of gender and age. *Sex Roles, 45*, 623–643.
Pike, J. J., & Jennings, N. A. (2005). The effects of commercials on children's perceptions of gender appropriate toy use. *Sex Roles, 52*, 83–91.
Pomerleau, A., Bolduc, D., Malcuit, G., & Cossette, L. (1990). Pink or blue: Environmental gender stereotypes in the first two years of life. *Sex Roles, 22*, 359–367.
Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Sciences, 4*, 66–72.
Pruden, S. M., & Abad, C. (2013). Do storybooks really break children's gender stereotypes?. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00986.
Reich, S. M., Black, R. W., & Foliaki, T. (2018). Constructing difference: LEGO set narratives promote stereotypic gender roles and play. *Sex Roles, 79*, 285–298.
Rhodes, M. (2012). Naïve theories of social groups. *Child Development, 83*, 1900–1916.
Rhodes, M., & Gelman, S. A. (2008). Categories influence predictions about individual consistency. *Child Development, 79*, 1270–1287.
Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 13526–13531.
Roberts, S. O., Ho, A. K., & Gelman, S. A. (2017). Group presence, category labels, and generic statements influence children to treat descriptive group regularities as prescriptive. *Journal of Experimental Child Psychology, 158*, 19–31.

Robinson, C. C., & Morris, J. T. (1986). The gender-stereotyped nature of Christmas toys received by 36-, 48-, and 60-month-old children: A comparison between nonrequested vs. requested toys. *Sex Roles, 15*, 21–32.
Ruble, D. N., Martin, C. L., & Berenbaum, S. A. (2007). Gender development. In S. Harter, W. Damon, & N. Eisenberg (Eds.), *Handbook of child psychology, Vol. 3: Social, emotional, and personality development* (pp. 858–932). New York: John Wiley.
Scott, K. P., & Feldman-Summers, S. (1979). Children's reactions to textbook stories in which females are portrayed in traditionally male roles. *Journal of Educational Psychology, 71*, 396–402.
Servin, A., Bohlin, G., & Berlin, L. (1999). Sex differences in 1-, 3-, and 5-year-olds' toy-choice in a structured play-session. *Scandinavian Journal of Psychology, 40*, 43–48.
Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General, 143*, 363–375.
Sherif, M. (1954). Integrating field work and laboratory in small group research. *American Sociological Review, 19*, 759–771.
Sherman, A. M., & Zurbriggen, E. L. (2014). "Boys can be anything": Effect of Barbie play on girls' career cognitions. *Sex Roles, 70*, 195–208.
Sherman, J. A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review, 74*, 290–299.
Shutts, K., Kenward, B., Falk, H., Ivegran, A., & Fawcett, C. (2017). Early preschool environments and gender: Effects of gender pedagogy in Sweden. *Journal of Experimental Child Psychology, 162*, 1–17.
Silver, I. M., & Shaw, A. (2018). Pint-sized public relations: The development of reputation management. *Trends in Cognitive Sciences, 22*, 277–279.
Snyder, T. D. (2018). Mobile digest of education statistics, 2017 (NCES No. 2018-138). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
Spinner, L., Cameron, L., & Calogero, R. (2018). Peer toy play as a gateway to children's gender flexibility: The effect of (counter) stereotypic portrayals of peers in children's magazines. *Sex Roles, 79*, 314–328.
Steinke, J., Lapinski, M. K., Crocker, N., Zietsman-Thomas, A., Williams, Y., Evergreen, S. H., & Kuchibhotla, S. (2007). Assessing media influences on middle school-aged children's perceptions of women in science using the Draw-A-Scientist Test (DAST). *Science Communication, 29*, 35–64.
Streiff, M., & Dundes, L. (2017). Frozen in time: How Disney gender-stereotypes its most powerful princess. *Social Sciences, 6*, 38.
Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology, 66*, 21–36.
Theimer, C. E., Killen, M., & Stangor, C. (2001). Young children's evaluations of exclusion in gender-stereotypic peer contexts. *Developmental Psychology, 37*, 18–27.
Todd, B. K., Fischer, R. A., Di Costa, S., Roestorf, A., Harbour, K., Hardiman, P., & Barry, J. A. (2018). Sex differences in children's toy preferences: A systematic review, meta-regression, and meta-analysis. *Infant and Child Development, 27*(2) e2064.
Van Breukelen, G. J. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*, 920–925.
Waxman, S. R. (2010). Names will never hurt me? Naming and the development of racial and gender categories in preschool-aged children. *European Journal of Social Psychology, 40*, 593–610.
Weeks, M. N., & Porter, E. P. (1983). A second look at the impact of nontraditional vocational role models and curriculum on the vocational role preferences of kindergarten children. *Journal of Vocational Behavior, 23*, 64–71.
Weisgram, E. S., & Bruun, S. T. (2018). Predictors of gender-typed toy purchases by prospective parents and mothers: The roles of childhood experiences and gender attitudes. *Sex Roles, 79*, 342–357.
Weisgram, E. S., Fulcher, M., & Dinella, L. M. (2014). Pink gives girls permission: Exploring the roles of explicit gender labels and gender-typed colors on preschool children's toy preferences. *Journal of Applied Developmental Psychology, 35*, 401–409.
Weisman, K., Johnson, M. V., & Shutts, K. (2015). Young children's automatic encoding of social categories. *Developmental Science, 18*, 1036–1043.
Witt, S. D. (1997). Parental influence on children's socialization to gender roles. *Adolescence, 32*, 253–259.
Zosuls, K. M., Ruble, D. N., Tamis-LeMonda, C. S., Shrout, P. E., Bornstein, M. H., & Greulich, F. K. (2009). The acquisition of gender labels in infancy: Implications for gender-typed play. *Developmental Psychology, 45*, 688–701.