CHILD DEVELOPMENT

Child Development, July/August 2020, Volume 91, Number 4, Pages e780–e798

Children's Appraisals of Gender Nonconformity: Developmental Pattern and Intervention

Karen Man Wa Kwan
University of Hong Kong

Sylvia Yun Shi
University of Hong Kong and The Chinese University of Hong Kong

A. Natisha Nabbijohn, Laura N. MacMullin, and Doug P. VanderLaan
University of Toronto Mississauga

Wang Ivy Wong
University of Hong Kong and The Chinese University of Hong Kong

Gender-nonconforming (GN) children are often perceived less positively, which may harm their well-being. We examined the development of such perceptions and an intervention to modify them. Chinese children's appraisals were assessed using multiple measures (verbal responses, sharing, and rank order task) after viewing vignettes of gender-conforming (GC) and GN hypothetical peers. In Study 1, children (N = 210; 4-, 5-, 8-, and 9-year-olds) were less positive toward GN than GC peers, especially if they were older or if the peers were boys. In Study 2 (N = 211, 8- and 9-year-olds), showing children exemplars of GN peers who displayed positive and GC characteristics subsequently reduced bias against gender nonconformity. These findings inform strategies aimed at reducing bias against gender nonconformity.

From an early age, children's lives are gendered. Most children can distinguish males and females by age two (Zosuls et al., 2009) and their knowledge of gender stereotypes increases with age (Signorella, Bigler, & Liben, 1993). Children prefer clothes, toys, and activities that are stereotypical to their gender (Maccoby, 1998; Wong & Yeung 2019), and play in gender-segregated groups (Martin et al., 2013). While these gender-related patterns appear to be robust developmental phenomena, individuals can also show gender nonconformity, the expression of cross-gender characteristics (Bailey & Zucker, 1995). Gender-nonconforming (GN) children break from stereotypical gender norms, such as those regarding appearance (e.g., clothing), behaviors (e.g., toy play), traits (e.g., nice) (Miller, Lurye, Zosuls, & Ruble, 2009), and gender of playmate preferences (Mehta & Strough, 2009). Depending on how gender nonconformity is defined and measured, the prevalence may vary. For example, approximately 1%–5% of children express marked gender nonconformity using two items from the Child Behavior Checklist about cross-gender behavior and wishes (van Beijsterveldt, Hudziak, & Boomsma, 2006), a more comprehensive scale, the Gender Identity Questionnaire for Children (van der Miesen, Nabbijohn, Santarossa, & VanderLaan, 2018), or a child-interview measure on perceived similarity between own gender and other gender (Martin, Andrews, England, Zosuls, & Ruble, 2017). Higher percentages have been reported when parents were asked to complete a checklist of everyday behaviors. In both the West and China, close to 20% of boys and 40% of girls of school age exhibited 10 or more different GN behaviors (Sandberg, Meyer-Bahlburg,

This research was funded by a Hong Kong Postgraduate Studentship for PhD study to Karen Man Wa Kwan, a Seed Fund for Basic Research of the University of Hong Kong and a Hong Kong Research Grants Council Early Career Scheme 27402914 to Wang Ivy Wong, a Social Sciences and Humanities Research Council (SSHRC) of Canada Graduate Scholarship-Master's Award to A. Natisha Nabbijohn, and a University of Toronto Connaught New Researcher Award and a SSHRC of Canada Insight Development Grant to Doug P. VanderLaan. We thank Nicole Raquinio for the artistic production of the stimuli, Diana Peragine and Alanna Santarossa for participation in study and stimuli design, and Icy Yeung Sui Ping for assistance in data collection.

Correspondence concerning this article should be addressed to Doug P. VanderLaan, Department of Psychology, University of Toronto, Rm. 4098, Deerfield Hall Building, 3359 Mississauga Rd, Mississauga, Ontario, L5L 1C6, Canada; or Wang Ivy Wong, Gender Studies Program, Faculty of Social Science, The Chinese University of Hong Kong, Rm. 249, Sino Building, Shatin, Hong Kong. Electronic mail may be sent to doug.vanderlaan@utoronto.ca or iwwong@cuhk.edu.hk.

DOI: 10.1111/cdev.13316

Ehrhardt, & Yager, 1993; Yu & Winter, 2011). Children tend to respond differently to peers who vary in their degree of gender nonconformity (Zucker, Wilson-Smith, Kurita, & Stern, 1995). Some degree of androgyny (i.e., presence of both gender-conforming [GC] and GN attributes) appears to be beneficial to adjustment (Martin et al., 2017). However, strongly GN children are likely to be less popular. Several studies indicated that children who express gender nonconformity are at risk of less positive peer relations (Aspenlieder, Buchanan, McDougall, & Sippola, 2009; Kowalski, 2007) as early as 3 years old (Langlois & Downs, 1980). Children are likely to reinforce gender norms in peer groups (Martin & Ruble, 2010), especially those of the same gender (Wallien, Veenstra, Kreukels, & Cohen-Kettenis, 2010). GN children tend to be evaluated less positively by peers than GC children (Carter & McCloskey, 1984; Kowalski, 2007; Langlois & Downs, 1980). Yet, little is known about the developmental pattern of children's responses to GN peers and what strategies, if any, might be effective for reducing children's bias against GN peers.

Children's Appraisals of Gender Nonconformity

Studies, often using verbal reports, found that children tend to have less positive appraisals of GN, than GC, peers (Blakemore, 2003; Carter & McCloskey, 1984; Levy, Taylor, & Gelman, 1995; Zucker et al., 1995). For example, in one study, although most elementary-school children thought fictional peers who violated gender norms were not wrong, they did not want to play with these peers (Carter & McCloskey, 1984). Levy et al. (1995) asked 4- and 8-year-old children and college students to report their feelings about GN boys and girls, whether they would like to be friends with them, and/or what they would say or do to them. This study also found negative responses toward gender norm violations.

Compared to GN girls, GN boys are especially likely to be rejected by peers (Braun & Davidson, 2017; Carter & McCloskey, 1984; Wallien et al., 2010). Leaper (1994) proposed that stereotypically masculine characteristics are perceived as being of higher status than stereotypically feminine characteristics and that higher status members tend to avoid characteristics associated with lower status members. This hypothesis may explain why it is more costly for boys, compared to girls, to violate gender norms.

It remains unclear whether children's appraisals of GN peers change as they age. A meta-analysis of children 2–13 years old found that gender stereotypes become more flexible with age (Signorella et al., 1993). Defining flexibility as the tendency to assign stereotypical attributes to both genders, Trautner et al. (2005) also found that children from 4.5 to 10 years old become more flexible with age. Flexibility is sometimes assessed by asking children about the ability or possibility for boys or girls to perform countergender-stereotypical activities (e.g., Can boys also play with dolls?). Compared to children in early childhood, children in middle childhood are more flexible in this regard (Blakemore, 2003; Carter & Patterson, 1982; Levy et al., 1995). However, increasing flexibility may mean greater awareness of the possibility of within-gender variations, but not necessarily more positivity in appraisal. For example, when asked whether they would like to play with or befriend GN peers, older children were less positive than younger children (Carter & McCloskey, 1984). Another study reported inconsistent associations between appraisals of gender nonconformity and age, but appraisals of most types of gender norm violations (7 out of 11) became more negative with age (Blakemore, 2003).

Thus, most studies have suggested a developmental decrease in positivity toward gender nonconformity from early to middle childhood, which would be consistent with the increase in gender-policing among children's peer groups with age. Children monitor and promote gender-norm adherence by laughing at peers who violate gender norms, teasing that they belong to the other-gender group, and telling them to correct their GN behaviors to GC behaviors (Kowalski, 2007). Peer pressure in same-gender play groups encourages children to maintain gender-stereotypical behaviors and/or punishes them for cross-gender behaviors (Lamb & Roopnarine, 1979; Langlois & Downs, 1980). For example, boys' groups tend to engage in more competitive and aggressive games, whereas girls' groups tend to have a more cooperative play style, and both boys and girls tend to look down on the activities of the other gender (Mehta & Strough, 2009). As a result, children tend to exhibit more gender-typed behaviors in same-gender peer groups (Martin & Fabes, 2001). With the dual effects of gender-policing and the peak of gender-segregation in middle childhood (Blakemore, 2003; Carter & McCloskey, 1984; Kowalski, 2007), older children may be less positive than younger children toward GN peers despite increased cognitive flexibility of gender norm assignments. In Study 1, we investigated how participants' age and gender, and the targets' gender and gender expression influenced children's appraisals of peers, broadly defined.

Interventions to Reduce Children's Bias against Gender Nonconformity

Although androgynous children may be well adjusted, strongly GN children are at risk of psychological maladjustment (Martin et al., 2017; van Beijsterveldt et al., 2006; van der Miesen et al., 2018; Zucker, Wood, & VanderLaan, 2014). Low peer acceptance may be the key risk factor. Clinical and community-based research of GN children has found that poor peer relations is a robust predictor of elevations in behavioral and emotional challenges among other factors such as social competence, IQ, parents' marital status, and social class (Cohen-Kettenis, Owen, Kaijser, Bradley, & Zucker, 2003; Kuvalanka, Weiner, Munroe, Goldberg, & Gardner, 2017; Roberts, Rosario, Slopen, Calzo, & Austin, 2013). Given the link between poor peer relations and lowered psychological well-being among GN children, it is worth examining possible intervention strategies. If children's bias against peers' gender nonconformity can be reduced, it may be possible to improve peer relations among GC and GN peers. In turn, these improved peer relations might help ameliorate lowered psychological well-being among GN children.

Some research has explored interventions aimed at reducing bias against gender nonconformity. For example, one study employed empathy, perspective taking and mere exposure in adolescents and young adults; however, it failed to increase positive attitudes toward sexual minorities and those who express gender nonconformity (Mundy-Shephard, 2015). Another study found that adults' attitudes toward GN children was more positive if the children possessed positive attributes, for example, being independent (GN girl) and being gentle (GN boy) (Coyle, Fulcher, & Trübutschek, 2016). However, this study examined adults' appraisals of children instead of children's appraisals of other children. Also, in Coyle et al. (2016), the description of the GN traits differed in valence (i.e., positive GN traits or negative GN traits) but there was no presentation of a GN girl with negative traits, limiting systematic comparisons of the effect of gender, gender nonconformity, and valence. Some studies (Lamb, Bigler, Liben, & Green, 2009; Pahlke, Bigler, & Martin, 2014) tested strategies aimed to get children to challenge sexist remarks/comments. These studies are relevant to fostering more positive appraisals of GN children, but they did not use scenarios in which target children were manipulated to differ on the degree of gender (non)conformity. Rather, they focused more on attitudes toward sexism in general (discrimination against a certain gender) than specifically on appraisals of GN children.

Interventions originally developed to reduce racial bias in children may provide a valuable starting point for devising approaches to reduce gender nonconformity bias because the roles of ethnicity and gender in the process of categorizing people into in-group and out-group members are thought to be similar. Ethnicity and gender are both perceptually salient features and children tend to focus on these features when categorizing people into in-group and out-group members (Bigler & Liben, 2007). Out-group members are usually associated with less positive traits (Bennett et al., 2004; Lam & Seaton, 2016). Thus, interventions that increase positive appraisals of racial out-groups may also increase positive appraisals of gender-based out-groups, such as GN peers.

Many interventions conducted in young children were ineffective in reducing racial bias (Aboud et al., 2012; Bigler, 1999), but presenting positive information about out-group members has yielded some success (Aboud & Doyle, 1996; Cameron, Rutland, Brown, & Douch, 2006; Hughes, Bigler, & Levy, 2007). For example, showing that racial minority children possess positive attributes such as diligence increased positive appraisals among children in the racial majority group toward the racial minority children (Litcher & Johnson, 1969). A recent intervention (presenting positive Black exemplars) improved implicit racial bias toward Black individuals in children 8–12 years old (Gonzalez, Steele, & Baron, 2016). In Study 2, we developed an intervention to improve children's appraisals of GN peers.

Study 1: Developmental Pattern

Aims

Study 1 examined how various aspects of children's appraisals of hypothetical target peers varied according to the participants' age and gender, and the targets' gender and gender expression. The target peers were either strongly GC (i.e., with only GC attributes) or strongly GN (i.e., with only GN attributes). There are many facets of peer appraisals. Prior studies relied on verbal reports to assess children's attitudes toward gender nonconformity. It is common to assess this by asking how much they would like to be friends with the target (Blakemore, 2003; Levy et al., 1995; Zucker et al., 1995). Some studies also assessed other aspects by asking children how happy the target was (Zucker et al., 1995), which is relevant to children's appraisal of the well-being of GN peers, and how much they would like to do what the target did in the story (Zucker et al., 1995), which provided insights about whether the children's liking of GN peers may be influenced by their own preference of performing that behavior, and whether they think what the target did in the story is wrong (Carter & McCloskey, 1984), which is directly relevant to children's appraisal of the moral component of peers' GN behavior. These aspects of peer appraisal were thus included in our study.

Verbal responses are limited by children's developing verbal skills and potentially also by social desirability bias (Paulhus, 1991). As such, in addition to verbal responses on rating scales, we used two other methods of assessment. First, sharing behavior is useful for measuring discrimination among children as it forces children to engage in a consequential "give-or-keep" decision. As a behavioral assessment, sharing is less explicit than verbal report. Also, sharing behavior has greater real-life relevance. Children in preschool already share more resources with family and friends than with strangers (Olson & Spelke, 2008), and more resources with friends than with non-friends (Moore, 2009). Children also tend to favor their own gender and own race in resource allocation (Renno & Shutts, 2015). These studies suggest that sharing is an important behavior often used by children to show friendliness. Second, ranking peers from most to least favorite is another useful measure that reflects children's social preference in real life. When children choose to interact with certain peers, doing so might often entail concomitantly foregoing time spent with other peers. Moreover, ranking is less susceptible to social desirability because children's answers cannot be equal for all targets. Thus, to measure appraisals of GC and GN targets more comprehensively, child participants responded verbally to a range of questions using rating scales, as well as by sharing stickers with the targets, and by ranking the targets in order from most to least favorite.

Hypotheses

To demonstrate that our sample showed gendered preferences typical of prior studies, we first tested the hypothesis that both boys and girls were more positive toward same-gender peers (*H_0*: Same-gender Peer Preference), a finding that has received strong empirical support in the West (Mehta & Strough, 2009). This hypothesis was tested in a preliminary analysis. Our main hypotheses were that children would be less positive toward GN than GC targets (Carter & McCloskey, 1984; Kowalski, 2007; Langlois & Downs, 1980; *H_1*: Less Positivity toward GN Peers), especially if the target was a boy (Braun & Davidson, 2017; Carter & McCloskey, 1984; Wallien et al., 2010; *H_2*: Male Bias). Also, we predicted that, compared to younger children, older children would be less positive toward GN targets (*H_3*: Older Children Harshness), because of peer gender-policing and increased gender-segregation in middle childhood (Kowalski, 2007; Lamb & Roopnarine, 1979; Langlois & Downs, 1980; Maccoby, 1998; Mehta & Strough, 2009).

We expected these hypotheses to hold in our Hong Kong Chinese sample because, despite variations in timing, magnitude, and specific content, the general patterns of gender development appear to be similar across industrialized cultures (Wong & VanderLaan, accepted). For example, the stages of gender identity and the broad content of gender stereotypes are largely consistent across cultures (Gibbons, 2000). A 3-year longitudinal study showed that middle-school boys in China and the United States showed similar adherence to gender-typed behaviors over time (Gupta et al., 2013). Also, like their U.S. counterparts, Chinese boys feel more pressure to conform to gender stereotypes than girls (Yu & Xie, 2010).

Method

Participants

Hong Kong Chinese children fluent in spoken Cantonese Chinese and aged 4, 5, 8, and 9 years old were recruited through kindergartens, primary schools and education centers, and through advertisements posted online and on campus. Interested parents were invited to fill in an online form. After screening for age and language requirement, 229 children were recruited from July to October 2017; however, 19 children were excluded because they were reported to have special educational needs, did not comply with the procedures of the study, did not meet the age requirement, or were from the same family as another participant (this last exclusion criterion was used to ensure the observations were independent). The final sample consisted of

210 participants, in which 104 children (52 boys and 52 girls) were in the younger age group (4–5 years old; $M_{\text{age}} = 4.98$, $SD = 0.49$), whereas 106 children (52 boys and 54 girls) were in the older age group (8–9 years old; $M_{\text{age}} = 8.99$, $SD = 0.56$). See Table 1 for demographic characteristics of the sample.

Procedures

Approval for this study was obtained from the Human Research Ethics Committee of a university in Hong Kong. Each child came to the laboratory with a parent. After giving consent, the parent completed a demographic questionnaire and other scales for a larger study. Meanwhile, the child was led by a researcher into another room. After giving verbal assent, they viewed vignettes about GC boy, GN boy, GC girl, and GN girl on a computer screen, after which they completed a verbal interview, a sticker sharing task, and a ranking task. They were told that this study was about children's perception of peer behaviors. The procedures took about 1 hr and participants received Hong Kong dollars 150, stationery, stickers, and a certificate as honoraria.

Older children (8- and 9-year-olds) in Study 1 also served as the control group of an intervention study, in which they viewed a slideshow about zoo animals at the beginning of the study (whereas an additionally recruited experimental group viewed positive exemplars of GN children; see details in Study 2). To maintain the consistency of the procedures between the older and younger children in Study 1, the younger children (4 and 5 years old) in Study 1 also viewed the same slideshow of zoo animals at the beginning of the study.

Materials

Vignettes of four targets. The main stimuli were standardized vignettes of four target children (GC boy, GN boy, GC girl, and GN girl) presented in random order. To rule out the potential effects of racial bias and any effect of the color of objects, all target children appeared Asian-looking and the use

Table 1
Demographic Characteristics of Participants

		Study 1		Study 2	
Variables	Range	*N*	*M* (*SD*)	*N*	*M* (*SD*)
Child age	4–9 (Study 1) 8–9 (Study 2)	210	6.52 (2.07)	211	8.50 (0.50)
Number of brothers	0–2	210	0.37 (0.54)	211	0.41 (0.59)
Number of sisters	0–3	210	0.41 (0.57)	211	0.44 (0.58)
Education level of the participating parent[a]	1–7	210	3.77 (1.75)	211	3.60 (1.82)
Education level of the participating parent's partner[a]	1–7	207	3.89 (1.76)	207	3.72 (1.92)
Income[b]	1–14	209	10.68 (2.11)	210	10.32 (2.56)
Family religious level[c]	0–5	210	1.05 (1.34)	211	1.14 (1.35)
Ethnicity	NA	210	Single-ethnic Chinese: 206 (98.10%) Multi-ethnic Chinese: 4 (1.90%)	211	Single-ethnic Chinese: 208 (98.58%) Multi-ethnic Chinese: 3 (0.01%)
Participating parent gender	NA	210	Female (mom): 185 Male (dad): 25	211	Female (mom): 180 Male (dad): 31
Family religion	NA	210	No religion: 121 Roman Catholic: 26 Protestant: 4 Christian: 51 Buddhist: 7 Hindu: 1	211	No religion: 111 Roman Catholic: 23 Protestant: 5 Christian: 56 Buddhist: 13 Hindu: 2 Other Religion: 1

[a] For education level, 1: Less than high school; 2: High school; 3: Diploma/Certificate; 4: Associate Degree; 5: University, Bachelor's degree; 6: University, Master's degree; 7: University, Doctoral degree. [b] For income (in Hong Kong dollars), the ratings stand for 1: < \$2,000; 2: \$2,000–\$3,999; 3: \$4,000–\$5,999; 4: \$6,000–\$7,999; 5: \$8,000–\$9,999; 6: \$10,000–\$14,999; 7: \$15,000–\$19,999; 8: \$20,000–\$24,999; 9: \$25,000–\$29,999; 10: \$30,000–\$39,999; 11: \$40,000–\$59,999; 12: \$60,000–\$79,999; 13: \$80,000–\$99,999; 14: ≥ \$100,000. [c] For family religious level, the ratings stand for 5: extremely; 4: very; 3: moderately; 2: slightly; 1: not at all; 0: no religion.

of gender-typed colors was avoided when the vignettes did not involve the targets' preferences with respect to clothing or toys.

Each vignette was 75 s long, consisting of five illustrations presented for 15-s each. Each illustration was accompanied by prerecorded audio narratives. The first illustration introduced the target's name and grade (the same grade as the participant). The other four illustrations described the target's preferences for toys, activities, clothing and hairstyle, and gender of playmates. Except for their names, the preferences of the GC boy and the GN girl were the same and those of the GC girl and the GN boy were also the same. The order of vignettes was randomized. Samples of vignettes with scripts can be found in Appendix S1.

Attention check. After each vignette, we tested whether the child had paid attention (e.g., "What is [Name of the target]'s favorite toy?"). If the child answered the question incorrectly, the illustration was repeated. The next illustration was shown only after the child gave a correct answer.

Children's responses. We measured children's responses toward the four targets: (a) Friendship preferences (Q_1: "Would you like being friends with [Name of the target]?"); (b) Perceived popularity (Q_2: "Do you think other children would like to be friends with [Name of the target]?"); (c) Emotion perception (Q_3: "Do you think [Name of the target] is happy?"); (d) Activity preferences (Q_4: "Would you like to do what [Name of the target] did in the story?"); (e) Moral judgment (Q_5: "Was what [Name of the target] was doing in the story right?"); (f) Sharing behavior (Q_6); and (g) Ranking (from most favorite target to least favorite target; Q_7). Children answered Q_1–Q_5 on a 3-point scale of 3 (*yes*), 2(*don't know*) and 1 (*no*) with corresponding emojis (i.e., smiling, neutral, frowning) as illustrations. Sharing behavior (Q_6) was measured by asking children to distribute ten stickers to the four targets and himself/herself in any manner they wished. To remind children of the four targets, pictures of the four targets and their preferences for toys, activities, clothing and hairstyle, and gender of playmates were shown when children allocated the stickers. Each child was allowed to change his/her decision once. To ensure that the child attended to and understood the instructions, the child was asked the total number of stickers s/he had before and after sharing. If the child failed to answer the question, the experimenter explained the instructions again until s/he understood. For ranking (Q_7), children were asked to rank the four targets in order from the most favorite target to the least favorite target. During this task, the pictures of the four targets were shown again. Responses expressed verbally and/or by pointing to the targets were accepted. All responses were coded in a way such that larger values in ratings, sticker sharing, and ranking indicated more positivity.

Results

Chi-square and *t*-tests were conducted to test for group differences in the demographic variables listed in Table 1 (see Table 1 for demographic characteristics of Study 1). There were no significant differences between the two age groups or genders in any demographic variable (all $ps > .05$). Therefore, these variables were not included as covariates.

Age and gender effects were analyzed in a series of 2 (child gender) × 2 (child age) × 2 (target gender) × 2 (target gender expression) repeated measures analyses of variance (ANOVAs), with Child Gender and Child Age as between-subjects factors and Target Gender and Target Gender Expression as within-subjects factors. The dependent variables were friendship preference (Q_1), perceived popularity (Q_2), emotion perception (Q_3), activity preferences (Q_4), moral judgment (Q_5), and sticker sharing (Q_6). For rank (Q_7), we used nonparametric tests including Friedman, Wilcoxon Signed-ranks, and Mann–Whitney *U* tests. For analyses on the three main hypotheses (H_1, H_2, H_3), when there was a four-way interaction, we followed up with further analyses to test the subsumed three-way interaction. If the three-way interaction was also significant, we then tested the subsumed two-way interaction. If a main effect or interaction was subsumed to a higher level interaction, we focused on the highest level interaction. Similar analytic approaches were used in prior research to follow-up higher order interactions (Taylor, Rhodes, & Gelman, 2009; Vlamings, Jonkman, & Kemner, 2010). Consequently, only the highest order effects are reported in text. All other significant effects are presented in Table S1. Table 2 summarizes the extent to which each hypothesis was supported. To reduce Type I error, all reported *p*-values were Bonferroni-adjusted for the number of questions asked (i.e., seven; for main and interaction effects) or for the number of tests needed to follow-up an interaction. However, for the preliminary analysis (H_0), we only focused on the planned Child Gender × Target Gender interaction because the aim was to test whether our sample showed a same-gender peer preference as would be expected from prior studies. The correlation matrices of the

Table 2
Summary of Hypothesis Testing Results in Study 1

Outcome variables	H_0: same-gender peer preference	H_1: less positivity toward GN peers	H_2: male bias	H_3: older children harshness
Friendship preference (Q_1)	Supported in female children	Supported in younger and older children rating same-gender targets	Supported in older male children	Supported in rating same-gender targets
Perceived popularity (Q_2)	NA	Supported	Not supported	Supported
Emotion perception (Q_3)	NA	Not supported	Not supported	Not supported
Activity preferences (Q_4)	NA	Supported in younger male children, and older children rating same-gender targets	Partially supported in younger male children and supported in older male children	Supported
Moral judgment (Q_5)	NA	Supported	Not supported	Supported
Sticker Sharing (Q_6)	Supported in male children	Supported in rating same-gender targets	Partially supported	Supported
Rank (Q_7)	Supported	Supported in rating same-gender targets	Supported	Supported in rating GN boy only

Note. GN = gender-nonconforming.

dependent variables for the four targets are provided in Table S3a.

We conducted power analyses in G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) for repeated measures analysis of variance with interactions among within- and between-subjects factors. Both Studies 1 and 2 had four groups and four repeated measures (i.e., one for each target child). With our lowest Bonferroni-corrected alpha of .007 and estimating for small-to-large correlations between repeated measures of r = .00–.70—the general range observed for the Study 1 and Study 2 samples (see Table S3)—our sample sizes (N = 210 in Study 1; N = 211 in Study 2) provided sufficient power to detect small and small-to-medium effects of approximately f = .11–.15 to f = .09–.17 (Cohen, 1969).

Preliminary Analysis (Replication of Same-Gender Preference)

The hypothesis (H_0) that children prefer same-gender peers was relevant to friendship preference, sticker sharing, and rank. As hypothesized, there were significant Child Gender × Target Gender interactions in friendship preference, $F(1, 206) = 11.47$, $p = .006$, $\eta_p^2 = .053$ and sticker sharing, $F(1, 206) = 15.19$, $p = .001$, $\eta_p^2 = .069$. Specifically, girls preferred same-gender ($M = 2.56$, $SD = 0.58$) to other-gender targets ($M = 2.35$, $SD = 0.69$) as friends, $t(105) = -3.38$, $p = .002$, $d = 0.33$, although boys did not. Boys shared more stickers with same-gender ($M = 1.76$, $SD = 0.72$) than other-gender targets ($M = 1.53$, $SD = 0.61$), $t(103) = 3.32$, $p = .003$, $d = 0.35$, although girls did not.

For rank, Wilcoxon Signed-ranks tests were conducted. For female participants, the mean rank favored same-gender targets ($M = 3.08$, $SD = 0.42$) over other-gender targets ($M = 1.92$, $SD = 0.43$), $Z = -7.94$, $p < .001$, $r = .77$. For male participants, the mean rank slightly favored other-gender targets ($M = 2.21$, $SD = 0.53$) over same-gender targets ($M = 2.79$, $SD = 0.53$), $Z = -4.57$, $p < .001$, $r = .45$. However, if the comparison focused on the GC targets, the female participants favored the GC girl ($M = 3.80$, $SD = 0.52$) over the GC boy ($M = 1.98$, $SD = 0.96$), $Z = -8.31$, $p < .001$, $r = .81$, whereas the male participants favored the GC boy ($M = 3.65$, $SD = 0.77$) over the GC girl ($M = 2.09$, $SD = 1.02$), $Z = -6.98$, $p < .001$, $r = .68$. Overall, the results supported the hypothesis that children prefer same-gender peers.

Friendship Preference (Q_1)

The highest order effect was a four-way Child Gender × Child Age × Target Gender × Target Gender Expression interaction, $F(1, 206) = 43.79$, $p < .001$, $\eta_p^2 = .175$. Post hoc analyses were first conducted within each age group to test H_1 (Less Positivity toward GN Peers) and H_2 (Male Bias).

In the younger group, boys and girls overall preferred being friends with GC targets ($M = 2.42$, $SD = 0.65$) to GN targets ($M = 2.28$, $SD = 0.72$), $d = 0.21$. Thus, H_1 (Less Positivity toward GN Peers) was supported. However, there was no Target Gender × Target Gender Expression interaction, so H_2 (Male Bias) was not supported.

In the older group, the boys preferred the GC ($M = 2.77$, $SD = 0.51$) to the GN boy ($M = 1.50$, $SD = 0.78$), $t(51) = 10.86$, $p < .001$, $d = 1.93$. However, they preferred the GN girl ($M = 2.56$, $SD = 0.70$) to the GC girl ($M = 1.71$, $SD = 0.87$), $t(51) = -6.12$, $p < .001$, $d = 1.07$. Thus, H_1 (Less Positivity toward GN Peers) was partially supported in older boys' friendship preference of same-gender targets. Moreover, the GN boy ($M = 1.50$, $SD = 0.78$) was less preferred than the GN girl ($M = 2.40$, $SD = 0.77$), $t(51) = -5.47$, $p < .001$, $d = 1.43$, which supported H_2 (Male Bias). The older girls preferred the GC girl ($M = 2.83$, $SD = 0.47$) to the GN girl ($M = 2.17$, $SD = 0.93$), $t(53) = 4.95$, $p < .001$, $d = 0.91$, but there was no such difference in their preference of male targets. However, H_2 (Male Bias) was not supported because older girls' preference for the GN girl and GN boy did not differ and they actually showed more bias toward gender nonconformity if the targets were girls than if the targets were boys.

To test H_3 (Older Children Harshness), we conducted further post hoc analyses on the four-way interaction within each child gender so that the age effect could be tested. Among the boys, older boys preferred the GC boy ($M = 2.77$, $SD = 0.51$) more than did younger boys ($M = 2.38$, $SD = 0.77$), $t(88.40) = -3.00$, $p = .007$, $d = 0.59$, unequal variance, and they preferred the GN boy ($M = 1.50$, $SD = 0.78$) less than did younger boys ($M = 2.12$, $SD = 0.88$), $t(102) = 3.78$, $p < .001$, $d = 0.74$. However, older boys preferred the GC girl ($M = 1.71$, $SD = 0.87$) less than did younger boys ($M = 2.17$, $SD = 0.76$), $t(100.16) = 2.88$, $p = .010$, $d = 0.56$, unequal variance, and preferred the GN girl ($M = 2.56$, $SD = 0.70$) more than did younger boys ($M = 2.08$, $SD = 0.86$), $t(97.86) = -3.13$, $p = .005$, $d = 0.61$, unequal variance. In sum, H_3 (Older Children Harshness) was partially supported in boys' friendship preference of same-gender targets. Among girls, older girls ($M = 2.17$, $SD = 0.93$) preferred the GN girl less than did younger girls ($M = 2.56$, $SD = 0.73$), $t(99.88) = 2.42$, $p = .034$, $d = 0.47$, unequal variance, but no age difference was found in their preference of male targets. Thus, H_3 (Older Children Harshness) was partially supported in girls' friendship preference of same-gender targets.

Perceived Popularity (Q_2)

The highest order effect was a Child Age × Target Gender Expression interaction, $F(1, 206) = 7.38$, $p = .05$, $\eta_p^2 = .035$. Post hoc comparisons showed that both younger and older children perceived GC targets ($M = 2.71$, $SD = 0.44$ and $M = 2.69$, $SD = 0.45$, respectively) as more popular than GN targets ($M = 2.56$, $SD = 0.59$ and $M = 2.33$, $SD = 0.61$, respectively), $t(103) = 3.50$, $p = .001$, $d = 0.29$ and $t(105) = 5.65$, $p < .001$, $d = 0.67$, respectively, consistent with H_1 (Less Positivity toward GN Peers). Moreover, alternative follow-up of the interaction showed that older children ($M = 2.33$, $SD = 0.61$) perceived GN targets as less popular than did younger children, ($M = 2.56$, $SD = 0.59$), $t(208) = 2.83$, $p = .010$, $d = 0.39$, whereas there was no significant age difference in rating for GC targets. Thus, H_3 (Older Children Harshness) was supported. However, because there was no interaction between Target Gender and Target Gender Expression, H_2 (Male Bias) was not supported.

Emotion Perception (Q_3)

There was a main effect of Child Age, $F(1, 206) = 10.06$, $p = .012$, $d = 0.43$, with older children ($M = 2.94$, $SD = 0.19$) perceiving the targets as happier than did younger children ($M = 2.81$, $SD = 0.39$). Apart from this effect, there were no other significant effects. Therefore, none of the hypotheses were supported.

Activity Preferences (Q_4)

The highest order effect was a four-way Child Gender × Child Age × Target Gender × Target Gender Expression interaction, $F(1, 206) = 30.27$, $p < .001$, $\eta_p^2 = .128$. Post hoc analyses were first conducted within each age group to test H_1 (Less Positivity toward GN Peers) and H_2 (Male Bias).

For younger children, boys preferred the activity of the GC boy ($M = 2.44$, $SD = 0.80$) to that of the GN boy ($M = 2.10$, $SD = 0.91$), $t(51) = 2.35$, $p = .046$, $d = 0.40$, whereas girls preferred the activity of the GN boy ($M = 2.60$, $SD = 0.75$) to that of GC boy ($M = 2.35$, $SD = 0.88$), $t(51) = -2.36$, $p = .044$, $d = 0.31$. Neither younger boys' nor younger girls' ratings for the GC Girl and the GN Girl differed significantly. Thus, H_1 (Less Positivity toward GN Peers) was partially supported in younger boys' activity preferences. Younger children's activity preferences for the GN targets did not differ by target gender. However, because boys showed less positivity toward GN when the target was a boy and not when the target was a girl, H_2 (Male Bias) was partially supported in younger boys.

For older children, boys preferred the activity of the GC boy ($M = 2.71$, $SD = 0.64$) to that of the GN boy ($M = 1.19$, $SD = 0.56$), $t(51) = 13.62$, $p < .001$, $d = 2.53$, but they preferred the activity of the GN girl ($M = 2.44$, $SD = 0.85$) to that of the GC girl ($M = 1.29$, $SD = 0.67$), $t(51) = -8.88$, $p < .001$, $d = 1.51$. For older girls, they preferred the activity of the GC girl ($M = 2.54$, $SD = 0.79$) to that of the GN girl ($M = 1.78$, $SD = 0.93$), $t(53) = 4.93$, $p < .001$, $d = 0.88$, although there was no difference for the target boys. Overall, H_1 (Less Positivity toward GN Peers) was supported in older children's activity preferences of same-gender targets. Moreover, older boys (but not girls) preferred the activity of the GN girl ($M = 2.44$, $SD = 0.85$) to that of the GN boy ($M = 1.19$, $SD = 0.56$), $t(51) = -9.73$, $p < .001$, $d = 1.74$, which partially supported H_2 (Male Bias).

In order to test the age effect predicted in H_3, we conducted further post hoc analyses for the four-way Child Gender × Child Age × Target Gender × Target Gender Expression interaction. The post hoc analyses were conducted separately for boys and girls.

For boys, for both the GN boy and the GC girl activities, younger boys ($M = 2.10$, $SD = 0.91$ and $M = 2.02$, $SD = 0.90$, respectively) preferred their activities more than did older boys ($M = 1.19$, $SD = 0.56$ and $M = 1.29$, $SD = 0.67$, respectively), $t(84.74) = 6.08$, $p < .001$, $d = 1.19$, and $t(94.21) = 4.72$, $p < .001$, $d = 0.93$, respectively (both unequal variances). Therefore, H_3 (Older Children Harshness) was partially supported in boys' activity preferences in response to target boys.

For girls, older girls preferred the activities of GC targets ($M = 2.28$, $SD = 0.68$) to those of GN targets ($M = 1.96$, $SD = 0.66$), $t(53) = 4.40$, $p < .001$, $d = 0.47$, but there was no such difference in younger girls. Thus, H_3 (Older Children Harshness) was supported in girls' activity preferences.

Moral Judgment (Q_5)

The highest order effect was a Child Age × Target Gender Expression interaction, $F(1, 206) = 13.96$, $p = .002$, $\eta_p^2 = .063$. Post hoc comparisons suggested that both younger and older children thought that the GC targets' ($M = 2.77$, $SD = 0.42$ and $M = 2.78$, $SD = 0.44$, respectively) behavior was more right than the GN targets' behavior ($M = 2.61$, $SD = 0.59$ and $M = 2.33$, $SD = 0.69$, respectively), though this tendency was weaker in the younger children, $t(103) = 3.80$, $p = .001$, $d = 0.31$, than in the older children, $t(105) = 6.84$, $p < .001$, $d = 0.78$. Moreover, older children were less likely to consider the activities of GN targets as right than were younger children, $t(203.90) = 3.23$, $p = .003$, $d = 0.45$, unequal variance, whereas there was no age difference in their moral judgments of the GC targets' activities. Therefore, both H_1 (Less Positivity toward GN Peers) and H_3 (Older Children Harshness) were supported. However, there was no significant Target Gender × Target Gender Expression interaction, so H_2 (Male Bias) was not supported.

Sticker Sharing (Q_6)

We found a two-way Child Age × Target Gender Expression interaction, $F(1, 206) = 22.77$, $p < .001$, $\eta_p^2 = .10$, and a three-way Child Gender × Target Gender × Target Gender Expression interaction, $F(1, 206) = 6.11$, $p < .001$, $\eta_p^2 = .133$. To test H_1 (Less Positivity toward GN Peers) and H_2 (Male Bias), post hoc analyses were first conducted to investigate the three-way interaction in boys and girls separately.

For boys, they shared fewer stickers with the GN boy ($M = 1.46$, $SD = 0.80$) than the GC boy ($M = 2.07$, $SD = 0.93$), $t(103) = 6.44$, $p < .001$, $d = 0.70$, but they shared similar numbers of stickers with the GC and GN girls. Thus, H_1 (Less Positivity toward GN Peers) was partially supported in boys' sharing with target boys. Boys' sharing with the GN boy and GN girl did not differ. However, they shared fewer stickers with the GN boy than the GC boy, whereas not showing similar discrimination toward the GN girl and GC girl, so H_2 (Male Bias) was supported.

For girls, they shared fewer stickers with the GN girl ($M = 1.39$, $SD = 0.61$) than the GC girl ($M = 1.90$, $SD = 0.86$), $t(105) = 5.99$, $p < .001$, $d = 0.68$, whereas there was no difference when the targets were boys. Therefore, H_1 (Less Positivity toward GN Peers) was partially supported in girls sharing with targets girls. However, girls' sharing discriminated against gender nonconformity when the targets were girls but not when the targets were boys, so H_2 (Male Bias) was not supported.

To test H_3 (Older Children Harshness), we further conducted post hoc comparisons of the Child Age × Target Gender Expression interaction. Older children shared more stickers with the GC targets ($M = 2.11$, $SD = 0.57$) than the GN targets ($M = 1.58$, $SD = 0.58$), $t(105) = 7.02$, $p < .001$. The younger children also shared more stickers with the GC targets ($M = 1.45$, $SD = 0.64$) than the GN targets ($M = 1.33$, $SD = 0.61$), $t(103) = 2.95$, $p = .008$; however, the difference in allocation in favor of GC targets was larger in the older children ($d = 0.93$) than in the younger children ($d = 0.19$). Therefore, H_3 (Older Children Harshness) was supported.

Rank (Q_7)

Friedman tests showed significant differences among the rankings of the GC boy, GC girl, GN boy, and GN girl both for male participants, $\chi^2(3, N = 104) = 116.18$, $p < .001$, and for female participants, $\chi^2(3, N = 106) = 152.74$, $p < .001$. Follow-up pairwise comparisons were conducted using Wilcoxon Signed-ranks tests. For male participants, the GC boy ($M = 3.65$, $SD = 0.77$) was ranked higher than the GN boy ($M = 1.92$, $SD = 0.91$), $Z = -7.80$, $p < .001$, $r = .77$, which supported H_1 (Less Positivity toward GN Peers). In addition, the GN girl ($M = 2.34$, $SD = 0.84$) was ranked higher than the GN boy ($M = 1.92$, $SD = 0.91$), $Z = -2.75$, $p = .024$, $r = .27$, which supported H_2 (Male Bias). For female participants, the GC girl ($M = 3.80$, $SD = 0.52$) was ranked higher than the GN girl ($M = 2.37$, $SD = 0.84$), $Z = -7.97$, $p < .001$, $r = .77$, which supported H_1 (Less Positivity toward GN Peers). In addition, the GN girl ($M = 2.37$, $SD = 0.84$) was ranked higher than the GN boy ($M = 1.86$, $SD = 0.84$), $Z = -3.42$, $p = .002$, $r = .33$, supporting H_2 (Male Bias).

Mann–Whitney *U* test was conducted to examine age group effects in the rankings of each of the target children. For the GN boy, younger children ($M = 2.12$, $SD = 0.88$) was ranked higher than that of the older children ($M = 1.67$, $SD = 0.81$), $U = 3{,}957.50$, $p < .001$, $r = .26$ but there was no age group difference in ranking for the other target children. Thus, older children were less positive than younger children toward the GN boy, which partially supported H_3 (Older Children Harshness).

Summary

The results of Study 1 (summarized in Tables 2 and S1) showed that children as young as 4–5 years of age gave less positive appraisals of and also shared less generously with peers who did not conform to stereotypical gender expressions. They were especially less positive if that peer was a boy, and the older children aged 8–9 years of age tended to be less positive toward GN target peers than did the younger children. These findings inform the developmental pattern of children's appraisals of gender nonconformity.

Study 2: Intervention to Reduce Gender Nonconformity Bias

Studies that have tested interventions aimed at reducing bias against gender nonconformity have focused on adult participants (Coyle et al., 2016; Mundy-Shephard, 2015), whereas relevant studies in children focused on sexism in general instead of gender nonconformity in particular (Lamb et al., 2009; Pahlke et al., 2014). In Study 2, we adapted an intervention from studies on racial bias reduction (Aboud & Doyle, 1996; Cameron et al., 2006; Hughes et al., 2007; Litcher & Johnson, 1969) to counteract children's less positive appraisals of GN peers. We predicted that child participants would respond more positively to GN targets when they were first presented with exemplars who displayed GC and socially desirable attributes (e.g., good academic performance) in addition to their GN attributes. We tested this intervention strategy in 8- to 9-year-old children. Study 1 and others (Blakemore, 2003; Carter & McCloskey, 1984; Levy et al., 1995) suggested that children at this age tend to be less positive than younger children toward gender norm violations. At the same time, older children show greater gender stereotype flexibility compared to younger children (Blakemore, 2003; Carter & Patterson, 1982; Levy et al., 1995). Rigidity in gender stereotype beliefs peaks at age 5–6 years, meaning that these beliefs become more flexible afterward, possibly as a result of more sophisticated cognitive development (Trautner et al., 2005). Older children become more able to attribute people's behaviors to their own personal choices or prerogatives, whereas younger children tend to rely on gender labels and gender expectations (Sinno & Killen, 2009). A similar intervention presenting positive Black exemplars to 5- to 12-year-old children also showed that positive change in implicit bias was successful in children aged 8 years old or above but not in those younger children (Gonzalez et al., 2016). Accordingly, we focused on the older children because when compared to younger children, older children's appraisals of GN peers may be more easily intervened by additional information about the targets' positive and GC attributes.

Method

Participants

Approval for this study was obtained together with Study 1 from the Human Research Ethnics Committee of a university in Hong Kong. The recruitment method was the same as in Study 1. Study 1 and Study 2 data were both collected from July to October 2017. All children completed the test phase (GC boy, GN boy, GC girl, and GN girl). All younger children and older children of Study 1 viewed the control (zoo) vignette first (the younger children did this just to ensure procedural equivalence). An additional group of older children viewed the intervention vignette (additional GN children with positive and GC attributes) first. Thus, Study 2 consisted of this additional group of older children (105 children: 53 boys and 52 girls; $M_{\text{age}} = 9.00$ and $SD = 0.56$) and the older children of Study 1 (106 children: 52 boys and 54 girls; $M_{\text{age}} = 8.99$ and $SD = 0.56$). The older children were randomly assigned to the control and intervention conditions. Initially, Study 2 included 220 Hong Kong Chinese children fluent in spoken Cantonese Chinese; however, nine children were excluded because they had special educational needs or they did not comply with the procedures of the study.

Experimental Condition

The intervention occurred before children viewed the vignettes of the four target children described in Study 1. In the experimental condition, children viewed a 3-min slideshow with audiovisual narratives. The first illustration introduced the names and grades of a boy and a girl (different from the targets described in Study 1). Then they were portrayed as violating gender expectations in toy play, activity, and clothing preferences (i.e., the boy likes to color books of his favorite Disney princesses and fairies, takes ballet lessons, and wears a pink princess dress; the girls likes to play with army men and Lego blocks, loves to wrestle, and dresses up like a soldier). However, they were also described as GC in some other aspects (i.e., the same boy enjoys basketball; the same girl loves to jump ropes), and as having some positive attributes (i.e., having lots of male and female friends, being good at catching caterpillars, and earning good grades at school). There were 12 illustrations in total, each lasting for 15 s. The order of slideshows was the same for all participants. See Appendix S2 for examples.

Control Condition

In the control condition, participants viewed a 3-min slideshow with audiovisual narratives describing zoo animals instead of children. This slideshow had an equal number of illustrations as the experimental condition, each lasting for 15 s. See Appendix S2 for examples.

Procedures

Children first viewed either the control or experimental slideshow. Then, they answered attention check questions after every three-to-four illustrations. Afterward, the remaining procedures were identical to those in Study 1.

Results

Chi-square tests and *t*-tests were conducted to test for any differences between the control and experimental groups for the demographic variables listed in Table 1. For all demographic variables, no group differences were found (all *p*s > .05), indicating that the randomization was successful with respect to these variables.

The intervention effect was analyzed in a series of 2 (condition: experimental vs. control) × 2 (child gender) × 2 (target gender) × 2 (target gender expression) ANOVAs, with Condition and Child Gender as between-subjects factors and Target Gender and Target Gender Expression as within-subjects factors. The dependent variables were friendship preference (Q_1), emotion perception (Q_3), activity preferences (Q_4), moral judgment (Q_5), sticker sharing (Q_6), and rank (Q_7). Perceived popularity (Q_2) was removed from analyses in Study 2 because it overlapped with one illustration (having lots of friends) in the intervention condition. For rank (Q_7), Mann–Whitney *U* tests were used. All *p*-values were Bonferroni-corrected and we used the same method as described in Study 1 to follow-up any higher order interactions. We only discuss the results pertaining to the intervention effect (i.e., effects involving Condition) in this study because the developmental effects free of the intervention were reported in Study 1. All the nonintervention-related findings of Study 2 are summarized in Table S2. The correlation matrices of the dependent variables for the four targets are provided in Table S3b.

Friendship Preference (Q_1)

There was a Condition × Child Gender × Target Gender × Target Gender Expression interaction, $F(1, 207) = 8.48$, $p = .024$, $\eta_p^2 = .039$. This four-way interaction was followed up by post hoc analyses within boys and girls, respectively.

For boys, the experimental group ($M = 1.96$, $SD = 0.85$) preferred the GN boy as friends more than the control group did ($M = 1.50$, $SD = 0.78$), $t(103) = -2.90$, $p = .009$, $d = 0.57$, although there was no difference between boys in the control versus experimental groups in how much they preferred the other three target children (see Figure 1a). No main effects or interactions were found when the target was female. For girls, no main effect or interaction was found. Thus, the hypothesized intervention effect was supported in boys' friendship preference concerning male targets.

Emotion Perception (Q_3)

There was no significant effect involving the intervention (see Figure 1b). The hypothesis was not supported.

Activity Preferences (Q_4)

There was a main effect of Condition, $F(1, 207) = 7.48$, $p = .041$, $d = 0.36$, and a Condition × Target Gender Expression interaction, $F(1, 207) = 8.09$, $p = .029$, $\eta_p^2 = .038$. Children's preferences for the activities of the GC targets did not differ by condition. However, the experimental group ($M = 2.19$, $SD = 0.59$) preferred the activities of the GN targets more than did the control group ($M = 1.89$, $SD = 0.61$), $t(209) = -3.55$, $p = .001$, $d = 0.49$ (see Figure 1c), which supported the hypothesis.

Moral Judgment (Q_5)

There was a Condition × Target Gender Expression interaction, $F(1, 207) = 10.91$, $p = .007$, $\eta_p^2 = .05$. No group differences were found for children's moral judgments about the GC targets. However, the experimental group ($M = 2.57$, $SD = 0.63$) was more likely to consider the activity of GN targets as right than was the control group ($M = 2.33$, $SD = 0.69$), $t(209) = -2.65$, $p = .017$, $d = 0.37$ (see Figure 1d), which supported the hypothesis.

Sticker Sharing (Q_6)

There was a Condition × Target Gender Expression interaction, $F(1, 207) = 14.39$, $p = .001$, $\eta_p^2 = .065$. The experimental group ($M = 1.91$, $SD = 0.48$) shared fewer stickers with the GC targets than did the control group ($M = 2.11$, $SD = 0.57$), $t(209) = 2.67$, $p = .016$, $d = 0.37$. Instead, the experimental group ($M = 1.75$, $SD = 0.52$) shared more stickers with the GN targets than did the control group ($M = 1.58$, $SD = 0.58$), $t(209) = -2.34$, $p = .040$, $d = 0.32$ (see Figure 1e), which supported the hypothesis.

Rank (Q_7)

The Mann–Whitney U test was significant in the averaged ranking of GC targets, $U = 3{,}770.50$, $p < .001$, $r = .30$, with the control group ($M = 3.03$, $SD = 0.44$) ranking them higher than the experimental group did ($M = 2.72$, $SD = 0.52$). The test was also significant in the averaged ranking of GN targets, with the experimental group ($M = 2.28$, $SD = 0.52$) ranking them higher than the control group did ($M = 1.97$, $SD = 0.44$), $U = 3{,}770.50$, $p < .001$, $r = .30$. Results of the Rank data, which supported the hypothesis, can be found in Figure 1f.

Summary

Our hypothesis was supported in most outcomes. Children in the experimental group were more positive toward GN targets when compared to children in the control group on activity preferences, moral judgment, sticker sharing, and ranking, and for male participants, friendship preference. The exception was emotion perception, which showed no effect of intervention.

General Discussion

Gender nonconformity is associated with less positive evaluations from peers (Blakemore, 2003; Carter & McCloskey, 1984; Levy et al., 1995; Zucker et al., 1995). However, little is known about the developmental pattern of these reactions, whether these reactions manifest in behaviors other than children's verbal reports using rating scales, and how to reduce such biases. This study extended findings to sharing behavior and ranking, and to an Asian sample. Study 1 examined children's attitudes toward GN peers and whether age and gender moderated the effect of gender expression. Study 2 broke new ground by demonstrating the effectiveness of a simple intervention in reducing bias toward gender nonconformity immediately posttest.

Less Positivity Toward Gender Nonconformity

By definition, most people tend to abide by gender norms; however, gender nonconformity is

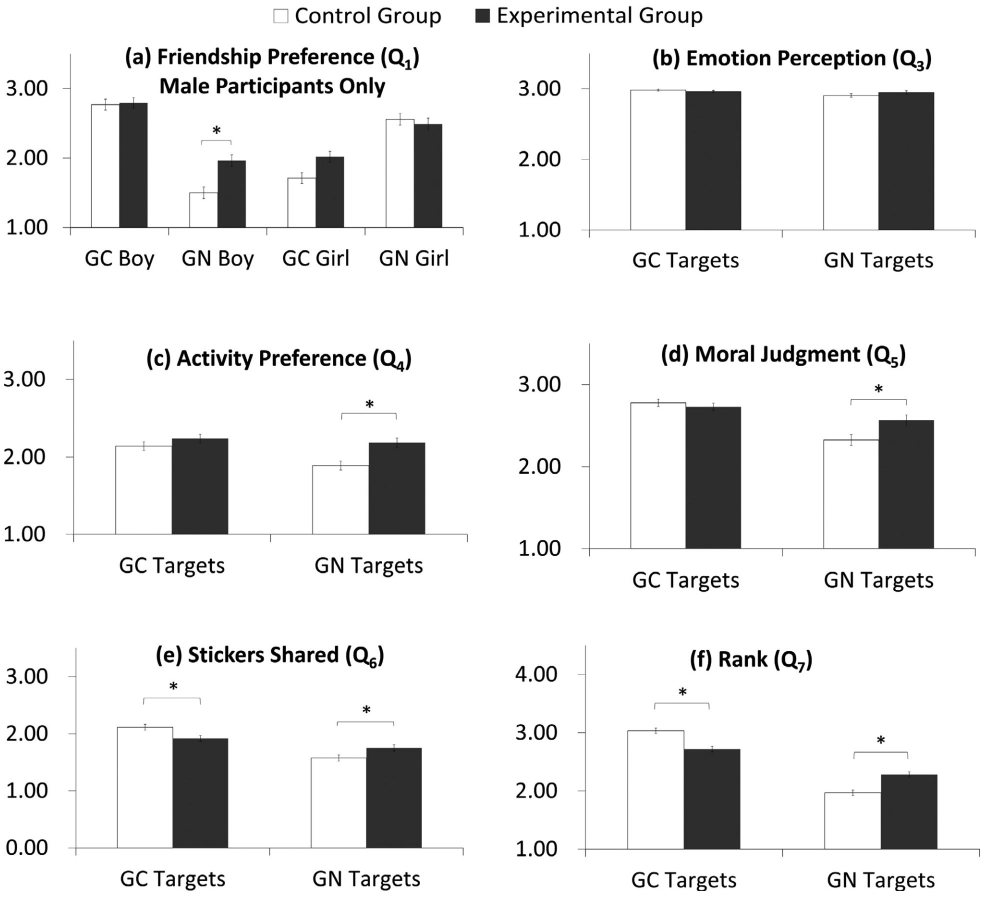

Figure 1. Mean ratings (A–D), number of stickers (E), and ranking (F) of control and experimental groups toward gender-conforming (GC) and gender-nonconforming (GN) targets. Larger values indicate more positivity. Perceived popularity (Q_2) was excluded from analysis in Study 2 because it overlapped with one of the manipulation illustrations (having lots of friends).

present in both clinical and nonclinical samples of children (Cohen-Kettenis et al., 2003; Martin et al., 2017; van Beijsterveldt et al., 2006; van der Miesen et al., 2018; Yu & Winter, 2011). It is estimated that around 20% of boys and 40% of girls of school age have shown ten or more different GN behaviors (Sandberg et al., 1993; Yu & Winter, 2011). Poor peer relations appears to place GN individuals at elevated risk of lowered psychological well-being (Cohen-Kettenis et al., 2003; Kuvalanka et al., 2017; Roberts et al., 2013). Therefore, it is important to study the developmental pattern of children's responses toward other children who are GN, especially given the latter constitute a substantial minority in the population.

In line with prior studies (Carter & McCloskey, 1984; Levy et al., 1995), Chinese children were less positive toward GN than GC hypothetical peers. Less positive evaluations were consistent, being statistically significant in 6 out of 7 outcomes. Of particular interest was that the less positive evaluations were not only present in verbal reports, but in sharing behavior and rank as well. These findings may demonstrate how rejection of gender nonconformity occurs in real life, given that sharing is used by children to show friendliness and to demonstrate closeness and favoritism (Moore, 2009; Olson & Spelke, 2008; Renno & Shutts, 2015). Also, in real life, prioritizing spending time with certain peers means foregoing the time one could spend with other peers.

It is also noteworthy that children's degree of preference for engaging in the activities depicted in the vignettes that were stereotypically associated with their own gender appeared to be influenced by whether the activity was being performed by a GC or GN target. Perhaps when children dislike certain peers, this disliking generalizes to the activities those peers perform. We found some indication that such may be the case. Children were generally less positive about GN than GC peers, and they tended to prefer the activities of same-gender peers when those activities aligned with those stereotypical for the participant's own gender. At the same time, however, the younger boys were not more positive about the activities performed by the GN girl than those by the GC girl, even though the activities of the GN girl were more stereotypically masculine and thus supposedly more attractive to boys. Similarly, the older girls were not more positive about the activities performed by the GN boy, compared to GC boy, even though the activities of the GN boy were more stereotypically feminine and supposedly more attractive to girls. As such, investigating the degree to which children's reduced positivity toward GN peers generalizes to other domains could be an important direction for future research.

The only exception to the reduced positivity was emotion perception—children did not think happiness had to do with one's gender expression. This finding is consistent with a previous finding that children rated both GC and GN targets as happy (Zucker et al., 1995). Perhaps this reflects that children (even at the younger ages of 4 and 5 years) think gender expression is a personal choice, similar to what others have found in older children (Sinno & Killen, 2009). Yet, other research suggests GN individuals are more likely to experience internalizing challenges (Martin et al., 2017; van Beijsterveldt et al., 2006; van der Miesen et al., 2018; Zucker et al., 2014). Thus, our finding could mean that children are not aware of the difficulties faced by GN peers.

Male Bias

In the West, GN boys are evaluated less positively than GN girls in peer nomination of classmates (Braun & Davidson, 2017; Wallien et al., 2010) and friendship preference of hypothetical children (Zucker et al., 1995). We found that children ranked the GN boy in a lower position than the GN girl. Less positivity toward the GN boy than the GN girl was also shown in the older (i.e., 8 and 9 years old) male participants' own friendship and activity preferences, the younger (i.e., 4 and 5 years old) male participants' activity preferences, and overall male participants' sharing behavior. Less positivity toward GN boys than girls may be explained by the higher status of masculinity than femininity as assigned by society. Higher status members are less likely to adopt the characteristics of lower status members, and society also tends to view it as less acceptable for boys to engage in cross-gender characteristics than girls (Leaper, 1994).

Older Children Harshness

Some prior studies suggested that children become less positive toward GN peers with age (Blakemore, 2003; Carter & McCloskey, 1984; Levy et al., 1995), but some studies found mixed patterns (Blakemore, 2003; Stoddart & Turiel, 1985). We found evidence that children become less positive, at least between 4–5 and 8–9 years of age, for 5 of 7 outcomes (perceived popularity, activity preferences, moral judgment, sticker sharing, and rank). Blakemore (2003) reported mixed patterns with 7 of 11 outcomes showing reduced positivity from youngest preschoolers to fifth graders and three showing a curvilinear pattern, in which the least positive evaluations were given by first and third graders as compared to younger and older children. The overall pattern of findings seems to converge on a decrease in positivity from early- to middle-childhood, with the possibility of an increase afterward.

The decrease from early- to middle-childhood in positivity toward GN peers may be related to same-gender peer preference, which emerges at around 2 years old and peaks in middle-childhood (Mehta & Strough, 2009). For example, when examining elementary children's social networks, cross-gender friends only account for 11% (Lee, Howes, & Chamberlain, 2007). Also, preadolescents expect more enjoyment with same-gender peers (Strough & Covatto, 2002). Gender segregation can contribute to more gender-typed behavior, which may relate to reduced positivity toward gender nonconformity (Martin & Fabes, 2001). Peers are important socialization agents in maintaining gendered behaviors by acting as gender police. For example, when recognizing other children's cross-gender behaviors, peers laugh at them and try to correct their behaviors (Kowalski, 2007).

Intervention

Without intervention, older children (8–9 years old) appraised GN peers less positively than younger children did, consistent with prior studies (Blakemore, 2003; Carter & McCloskey, 1984; Levy et al., 1995). Study 2 showed that the intervention was effective in the older children, significant in five of six outcomes (activity preferences, moral judgment, sticker sharing, ranking, and, for male participants, friendship preference), at least immediately following the intervention. By simply presenting target children with a diverse range of traits (both conforming and nonconforming, and traits that would be considered positive such as performing well in school), children were more positive toward GN hypothetical peers.

What mechanisms might explain the intervention effect? Individuals tend to categorize and generalize based on initial information. When forming impressions of others, stereotypic category-attribute associations are readily used (Fiske & Neuberg, 1990). Out-group members tend to be assigned with less positive characteristics (Bennett et al., 2004; Lam & Seaton, 2016) and viewed as more homogeneous (Brewer, 1993) than in-group members. It is possible that most children view GN peers as out-group members and thus form less positive appraisals toward them.

A recent study on racial bias by Gonzalez et al. (2016) employed an intervention similar to ours. By presenting positive Black exemplars, children aged 8–12 years old became more positive in their appraisals of Black individuals. Two explanations were proposed. First, presenting positive exemplars may alter the social context by focusing on individuals who differ from or challenge the usual negative stereotypes (Lai, Hoffman, & Nosek, 2013). This may prime subtypes of individuals (i.e., GN individuals with positive and GC attributes in this study). Another possibility is that by presenting GN targets as having a wide range of attributes (positive and GC attributes) instead of having homogenous GN attributes only, children's existing beliefs about the group may change by forming new associations about the GN peers that are more positive. In other words, one could speculate that this intervention encourages children to develop more positive appraisals of GN peers by priming a more positive out-group subcategory and/or by modifying existing categories.

In addition to these two previously proposed explanations, past research suggests that similarity promotes liking (Gilovich, Keltner, Chen, & Nisbett, 2013). Thus, it is also possible that this intervention worked by highlighting the shared attributes (positive and GC attributes) of the participants and the GN peers and, therefore, increased the perceived similarity between them. This study was not designed to discern which particular mechanism made the intervention effective, and future studies are needed to test possible mechanisms. Also, different mechanisms might apply to different groups of children. For example, the similarity-promotes-liking mechanism may be more relevant to children who are themselves GC (who, by definition, constitute the majority) than to children who are not, because they are the ones who are most likely to, without intervention, see themselves as being dissimilar from GN children.

It is important to note, however, that although target peers are often portrayed as either GC or GN in experimental research studies such as the present ones, individuals are rarely defined solely by gender expression and that peers are seldom GC or GN in all respects (Miller et al., 2009). Individuals often possess a myriad of attributes, some GC, some GN, and some negative, some positive. Our intervention method incorporated some of these more realistic features and tested the potential of a strategy that may be applicable in real life. Although we presented positive attributes (e.g., doing well as school) and GC attributes of the GN children as part of the manipulation, highlighting positive attributes of individuals and qualities that GC and GN children share more broadly (without highlighting whether they are GN or GC (e.g., such as how they both share in the identity of "student" at school) could be helpful. For example, having teachers create opportunities for GC and GN children to learn about how each person is good as an individual and ways that they are potentially similar would be worthwhile.

Limitations and Future Directions

This study is the first to use multiple measures to assess children's appraisals of GC versus GN peers in a controlled experimental design and to develop and test a possible intervention strategy for reducing children's bias against GN peers. Although the results showed developmental effects and supported the effectiveness of the intervention, there were a number of limitations that are important to note. First, the present research was conducted in a laboratory setting and used standardized hypothetical peers as targets to increase experimental rigor. This allowed for control of confounds but also reduced the ecological validity of the study. It would be valuable to study children's responses and the effectiveness of the intervention in naturalistic settings and using real peer targets by adopting a sociometric approach.

We used attention check questions for the vignettes to ensure that children attended to the information presented in all conditions and all vignettes. Thus, participants demonstrated that they remembered all the contents at least immediately following presentation of stimuli. However, that they might have forgotten about some of the manipulation content by the test phase, and it could be that certain attributes (e.g., GC attributes) were more easily forgotten than others (e.g., positive attributes). We cannot be certain whether the GC attributes, the positive attributes, or both, had an effect on increasing positive appraisals of the GN targets in the intervention. Future research could directly compare the effectiveness of each manipulation.

The attention check questions did not involve asking children the gender of the targets. Although we purposefully chose highly gendered names for the targets and the targets were shown with gender-typed hair (except in the illustration of clothing and hairstyle), there is still a chance that participants might have distorted the gender of the targets, especially when the targets were GN (Bigler & Liben, 1993). Future studies should consider adding the gender of targets in the attention check questions. Also, even though the targets in this study were all portrayed by the same graphic artist and were stylistically similar, the attractiveness of the targets was not examined explicitly and could be a confound. To reduce test demands, only one illustration was included for each domain (i.e., toy, activity, appearance, and gender of playmates); however, this can limit the generalizability of the findings.

In everyday life, unlike in the lab, the idea that GN individuals possess a myriad of attributes may not be emphasized as explicitly. This difference might explain why a relatively simple intervention was effective in our study, but discrimination against gender nonconformity is still pervasive in everyday life. Also, our intervention did not employ a longitudinal design. Rather, it focused on short-term effects and, thus, it is unclear whether our intervention would be effective in the longer term. Future research should discern whether the intervention can be effective in the longer term, and also whether it can be augmented in some ways to make it more effective long-term. The intervention used here was passive in its approach and its effectiveness could potentially be improved by incorporating an active training component. Some research has shown that, compared to a more passive approach, an active training approach increases the short- and long-term effectiveness of interventions that encourage children to challenge instances of sexism (Lamb et al., 2009). Our findings showed that younger children aged 4–5 years old held some bias toward GN children. Whether this intervention is effective in age groups other than those 8–9 year olds should also be explored in future studies. It is also relevant to note that the extent to which GN children would benefit from the current intervention may depend on the extent of their gender nonconformity, as peer appraisals may become increasingly negative as targets' level of gender nonconformity increases (Zucker et al., 1995).

Conclusion

This study provided useful insights into the developmental pattern of children's appraisals of GN peers and extended the current literature by employing sharing behavior, a more implicit measure, and rank, which assessed children's social preference, in measuring such appraisals. It also provided the first evidence that presenting GN behaviors alongside GC and positive attributes effectively reduced children's bias against GN peers. These findings provide valuable insights that may be useful for developing strategies aimed at ameliorating the stigma and discrimination that appear to place GN children at risk for poorer psychological well-being.

References

Aboud, F. E., & Doyle, A. B. (1996). Does talk of race foster prejudice or tolerance in children? *Canadian Journal of Behavioural Science*, *28*, 161. https://doi.org/10.1037/0008-400X.28.3.161

Aboud, F. E., Tredoux, C., Tropp, L. R., Brown, C. S., Niens, U., & Noor, N. M. (2012). Interventions to reduce prejudice and enhance inclusion and respect for ethnic differences in early childhood: A systematic review. *Developmental Review*, *32*, 307–336. https://doi.org/10.1016/j.dr.2012.05.001

Aspenlieder, L., Buchanan, C. M., McDougall, P., & Sippola, L. K. (2009). Gender nonconformity and peer victimization in pre-and early adolescence. *European Journal of Developmental Science*, *3*, 3–16. https://doi.org/10.3233/DEV-2009-3103

Bailey, J. M., & Zucker, K. J. (1995). Childhood sex-typed behavior and sexual orientation: A conceptual analysis and quantitative review. *Developmental Psychology*, *31*, 43–55. https://doi.org/10.1037//0012-1649.31.1.43

Bennett, M., Barrett, M., Karakozov, R., Kipiani, G., Lyons, E., Pavlenko, V., & Riazanova, T. (2004). Young children's evaluations of the ingroup and of outgroups: A multi-national study. *Social Development*, *13*, 124–141. https://doi.org/10.1046/j.1467-9507.2004.00260.x

Bigler, R. S. (1999). The use of multicultural curricula and materials to counter racism in children. *Journal of Social Issues*, *55*, 687–705. https://doi.org/10.1111/0022-4537.00142

Bigler, R. S., & Liben, L. S. (1993). A cognitive-developmental approach to racial stereotyping and reconstructive memory in Euro-American children. *Child Development*, *64*, 1507–1518. https://doi.org/10.1111/j.1467-8624.1993.tb02967.x

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. https://doi.org/10.1111/j.1467-8721.2007.00496.x

Blakemore, J. E. O. (2003). Children's beliefs about violating gender norms: Boys shouldn't look like girls, and girls shouldn't act like boys. *Sex Roles*, *48*, 411–419. https://doi.org/10.1023/A:1023574427720

Braun, S. S., & Davidson, A. J. (2017). Gender (non) conformity in middle childhood: a mixed methods approach to understanding gender-typed behavior, friendship, and peer preference. *Sex Roles*, *77*, 16–29. https://doi.org/10.1007/s11199-016-0693-z

Brewer, M. B. (1993). Social identity, distinctiveness, and in-group homogeneity. *Social Cognition*, *11*, 150–164. https://doi.org/10.1521/soco.1993.11.1.150

Cameron, L., Rutland, A., Brown, R., & Douch, R. (2006). Changing children's intergroup attitudes toward refugees: Testing different models of extended contact. *Child Development*, *77*, 1208–1219. https://doi.org/10.1111/j.1467-8624.2006.00929.x

Carter, D. B., & McCloskey, L. A. (1984). Peers and the maintenance of sex-typed behavior: The development of children's conceptions of cross-gender behavior in their peers. *Social Cognition*, *2*, 294–314. https://doi.org/10.1521/soco.1984.2.4.294

Carter, D. B., & Patterson, C. J. (1982). Sex roles as social conventions: The development of children's conceptions of sex-role stereotypes. *Developmental Psychology*, *18*, 812–824. https://doi.org/10.1037//0012-1649.18.6.812

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen-Kettenis, P. T., Owen, A., Kaijser, V. G., Bradley, S. J., & Zucker, K. J. (2003). Demographic characteristics, social competence, and behavior problems in children with gender identity disorder: A cross-national, cross-clinic comparative analysis. *Journal of Abnormal Child Psychology*, *31*, 41–53. https://doi.org/10.1023/A:1021769215342

Coyle, E. F., Fulcher, M., & Trübutschek, D. (2016). Sissies, mama's Boys, and tomboys: Is children's gender nonconformity more acceptable when nonconforming traits are positive? *Archives of Sexual Behavior*, *45*, 1827–1838. https://doi.org/10.1007/s10508-016-0695-5

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Amsterdam, The Netherlands: Elsevier.

Gibbons, J. L. (2000). *Gender development in cross-cultural perspective*. Hillsdale, NJ: Erlbaum.

Gilovich, T., Keltner, D., Chen, S., & Nisbett, R. E. (2013). *Social psychology* (3rd ed.). New York, NY: W.W. Norton.

Gonzalez, A. M., Steele, J. R., & Baron, A. S. (2016). Reducing children's implicit racial bias through exposure to positive out-group exemplars. *Child Development*, *88*, 123–130. https://doi.org/10.1111/cdev.12582

Gupta, T., Way, N., McGill, R. K., Hughes, D., Santos, C., Jia, Y., . . . Deng, H. (2013). Gender-typed behaviors in friendships and well-being: A cross-cultural study of Chinese and American boys. *Journal of Research on Adolescence*, *23*, 57–68. https://doi.org/10.1111/j.1532-7795.2012.00824.x

Hughes, J. M., Bigler, R. S., & Levy, S. R. (2007). Consequences of learning about historical racism among European American and African American children. *Child Development*, *78*, 1689–1705. https://doi.org/10.1111/j.1467-8624.2007.01096.x

Kowalski, K. (2007). The development of social identity and intergroup attitudes in young children. In O. Saracho & B. Spodek (Eds.), *Contemporary perspectives on social learning in early childhood education* (pp. 51–84). Charlotte, NC: IAP.

Kuvalanka, K. A., Weiner, J. L., Munroe, C., Goldberg, A. E., & Gardner, M. (2017). Trans and gender-nonconforming children and their caregivers: Gender presentations, peer relations, and well-being at baseline. *Journal of Family Psychology*, *31*, 889. https://doi.org/10.1037/fam0000338

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*, 315–330. https://doi.org/10.1111/spc3.12023

Lam, V. L., & Seaton, J. (2016). Ingroup/outgroup attitudes and group evaluations: The role of competition in British classroom settings. *Child Development Research*, *2016*, https://doi.org/10.1155/2016/8649132

Lamb, L. M., Bigler, R. S., Liben, L. S., & Green, V. A. (2009). Teaching children to confront peers' sexist remarks: Implications for theories of gender development and educational practice. *Sex Roles*, *61*, 361–382. https://doi.org/10.1007/s11199-009-9634-4

Lamb, M. E., & Roopnarine, J. L. (1979). Peer influences on sex-role development in preschoolers. *Child Development*, *50*, 1219–1222. https://doi.org/10.2307/1129353

Langlois, J. H., & Downs, A. C. (1980). Mothers, fathers, and peers as socialization agents of sex-typed play behaviors in young children. *Child Development*, *51*, 1237–1247. https://doi.org/10.2307/1129566

Leaper, C. (1994). Exploring the consequences of gender segregation on social relationships. *New Directions for Child and Adolescent Development*, *1994*, 67–86. https://doi.org/10.1002/cd.23219946507

Lee, L., Howes, C., & Chamberlain, B. (2007). Ethnic heterogeneity of social networks and cross-ethnic friendships of elementary school boys and girls. *Merrill-Palmer Quarterly*, *53*, 325–346. https://doi.org/10.1353/mpq.2007.0016

Levy, G. D., Taylor, M. G., & Gelman, S. A. (1995). Traditional and evaluative aspects of flexibility in gender roles, social conventions, moral rules, and physical laws. *Child Development*, *66*, 515–531. https://doi.org/10.2307/1131594

Litcher, J. H., & Johnson, D. W. (1969). Changes in attitudes toward Negroes of white elementary school students after use of multiethnic readers. *Journal of Educational Psychology*, *60*, 148–152. https://doi.org/10.1037/h0027081

Maccoby, E. E. (1998). *The two sexes: Growing up apart, coming together* (Vol. 4). Cambridge, MA: Harvard University Press.

Martin, C. L., Andrews, N. C., England, D. E., Zosuls, K., & Ruble, D. N. (2017). A dual identity approach for conceptualizing and measuring children's gender identity. *Child Development*, *88*, 167–182. https://doi.org/10.1111/cdev.12568

Martin, C. L., & Fabes, R. A. (2001). The stability and consequences of young children's same-sex peer interactions. *Developmental Psychology*, *37*, 431–446. https://doi.org/10.1037/0012-1649.37.3.431

Martin, C. L., Kornienko, O., Schaefer, D. R., Hanish, L. D., Fabes, R. A., & Goble, P. (2013). The role of sex of peers and gender-typed activities in young children's peer affiliative networks: A longitudinal analysis of selection and influence. *Child Development*, *84*, 921–937. https://doi.org/10.1111/cdev.12032

Martin, C. L., & Ruble, D. N. (2010). Patterns of gender development. *Annual Review of Psychology*, *61*, 353–381. https://doi.org/10.1146/annurev.psych.093008.100511

Mehta, C. M., & Strough, J. (2009). Sex segregation in friendships and normative contexts across the life span. *Developmental Review*, *29*, 201–220. https://doi.org/10.1016/j.dr.2009.06.001

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*, 870–881. https://doi.org/10.1007/s11199-009-9584-x

Moore, C. (2009). Fairness in children's resource allocation depends on the recipient. *Psychological Science*, *20*, 944–948. https://doi.org/10.1111/j.1467-9280.2009.02378.x

Mundy-Shephard, A. M. (2015). *Empathy, perspective-taking and the mere exposure effect: Understanding adolescent attitudes about sexual minorities and reducing prejudice against sexual minority youth*. Doctoral dissertation, Harvard Graduate School of Education.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*, 222–231. https://doi.org/10.1016/j.cognition.2007.12.003

Pahlke, E., Bigler, R. S., & Martin, C. L. (2014). Can fostering children's ability to challenge sexism improve critical analysis, internalization, and enactment of inclusive, egalitarian peer relationships? *Journal of Social Issues*, *70*, 115–133. https://doi.org/10.1111/josi.12050

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Renno, M. P., & Shutts, K. (2015). Children’s social category-based giving and its correlates: Expectations and preferences. *Developmental Psychology*, *51*, 533–543. https://doi.org/10.1037/a0038819

Roberts, A. L., Rosario, M., Slopen, N., Calzo, J. P., & Austin, S. B. (2013). Childhood gender nonconformity, bullying victimization, and depressive symptoms across adolescence and early adulthood: An 11-year longitudinal study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*, 143–152. https://doi.org/10.1016/j.jaac.2012.11.006

Sandberg, D. E., Meyer-Bahlburg, H. F., Ehrhardt, A. A., & Yager, T. J. (1993). The prevalence of gender-atypical behavior in elementary school children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *32*, 306–314. https://doi.org/10.1097/00004583-199303000-00010

Signorella, M. L., Bigler, R. S., & Liben, L. S. (1993). Developmental differences in children's gender schemata about others: A meta-analytic review. *Developmental Review*, *13*, 147–183. https://doi.org/10.1006/drev.1993.1007

Sinno, S. M., & Killen, M. (2009). Moms at work and dads at home: Children's evaluations of parental roles. *Applied Developmental Science*, *13*, 16–29. https://doi.org/10.1080/10888690802606735

Stoddart, T., & Turiel, E. (1985). Children's concepts of cross-gender activities. *Child Development*, *56*, 1241–1252. https://doi.org/10.2307/1130239

Strough, J., & Covatto, A. M. (2002). Context and age differences in same-and other-gender peer preferences. *Social Development*, *11*, 346–361. https://doi.org/10.1111/1467-9507.00204

Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; Cows will be cows: Children’s essentialist reasoning about gender categories and animal species. *Child Development*, *80*, 461–481. https://doi.org/10.1111/j.1467-8624.2009.01272.x

Trautner, H. M., Ruble, D. N., Cyphers, L., Kirsten, B., Behrendt, R., & Hartmann, P. (2005). Rigidity and flexibility of gender stereotypes in childhood: Developmental or differential? *Infant and Child Development*, *14*, 365–381. https://doi.org/10.1002/icd.399

van Beijsterveldt, C., Hudziak, J. J., & Boomsma, D. I. (2006). Genetic and environmental influences on cross-gender behavior and relation to behavior problems: A study of Dutch twins at ages 7 and 10 years. *Archives of Sexual Behavior*, *35*, 647–658. https://doi.org/10.1007/s10508-006-9072-0

van der Miesen, A. I., Nabbijohn, A. N., Santarossa, A., & VanderLaan, D. P. (2018). Behavioral and emotional problems in gender-nonconforming children: A Canadian community-based study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*, 491–499. https://doi.org/10.1016/j.jaac.2018.03.015

Vlamings, P. H., Jonkman, L. M., & Kemner, C. (2010). An eye for detail: An event-related potential study of the rapid processing of fearful facial expressions in children. *Child Development*, *81*, 1304–1319. https://doi.org/10.1111/j.1467-8624.2010.01470.x

Wallien, M. S., Veenstra, R., Kreukels, B. P., & Cohen-Kettenis, P. T. (2010). Peer group status of gender dysphoric children: A sociometric study. *Archives of Sexual Behavior*, *39*, 553–560. https://doi.org/10.1007/s10508-009-9517-3

Wong, W. I., & VanderLaan, D. P. (accepted). Early sex differences and similarities: Evidence across cultures? In F. M. Cheung & D. F. Halpern (Eds.), *Cambridge international handbook on psychology of women*. Cambridge, UK: Cambridge University Press.

Wong, W. I., & Yeung, S. P. (2019). Preschool gender differences in spatial and social skills and their relations to play and parental socialization in Hong Kong Chinese children. *Archives of Sexual Behavior*. https://doi.org/10.1007/s10508-019-1415-8

Yu, L., & Winter, S. (2011). Gender atypical behavior in Chinese school-aged children: Its prevalence and relation to sex, age, and only child status. *Journal of Sex Research*, *48*, 334–348. https://doi.org/10.1080/00224491003774867

Yu, L., & Xie, D. (2010). Multidimensional gender identity and psychological adjustment in middle childhood: A study in China. *Sex Roles*, *62*, 100–113. https://doi.org/10.1007/s11199-009-9709-2

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. https://doi.org/10.1037/a0014053

Zucker, K. J., Wilson-Smith, D. N., Kurita, J. A., & Stern, A. (1995). Children's appraisals of sex-typed behavior in their peers. *Sex Roles*, *33*, 703–725. https://doi.org/10.1007/bf01544775

Zucker, K. J., Wood, H., & VanderLaan, D. P. (2014). Models of psychopathology in children and adolescents with gender dysphoria. In B. Kreukels, T. Steensma, & A. de Vries (Eds.), *Gender dysphoria and disorders of sex development. Focus on Sexuality Research* (pp. 171–192). Boston, MA: Springer.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher’s website:

Table S1. Results for the Analyses of Variance (Q_1–Q_6) and Nonparametric Tests (Q_7) in Study 1

Table S2. *Nonintervention-Related Findings (i.e., Effects Not Involving Condition) in Study 2*

Table S3. (a) Study 1; (b) Study 2

Appendix S1. Description of Vignettes Presented to Test the Developmental Pattern of Appraisals of Gender Nonconformity

Appendix S2. Description of Vignettes Presented to Test the Intervention Effects