Testing an Empathy-Based Intervention to Improve Children's Gender-Related Attitudes

MacMullin, L. N.[1,2]
Wong, W. I.[3,4]
VanderLaan, D. P.[1,5]
[1] **Department of Psychology, University of Toronto Mississauga, Mississauga, Canada**
[2] **Department of Psychology, Middlebury College, Middlebury, United States**
[3] **Gender Studies Programme, The Chinese University of Hong Kong, Hong Kong**
[4] **Department of Psychology, The Chinese University of Hong Kong, Hong Kong**
[5] **Child and Youth Psychiatry, Centre for Addiction and Mental Health, Toronto, Canada**

Despite past research showing that children hold less positive attitudes toward gender-nonconforming (vs. gender-conforming) children, few studies have evaluated possible approaches to limit biases. Previous literature suggests empathy-based interventions can improve intergroup attitudes and reduce stereotyping. Thus, we designed a novel empathy-based intervention aimed at reducing children's gender stereotyping and improving their attitudes toward gender-nonconforming peers. We tested our intervention among 186 (51% girls, 48% control condition) 8- to 11-year-old Canadian children. The intervention was not effective at reducing children's gender stereotyping or their bias against gender-nonconforming peers. In girls, children's trait empathy was positively related to their attitudes toward target children. The Discussion outlines possible reasons why the intervention was ineffective and considers how future work can be strengthened.

Keywords: empathy intervention, gender nonconformity, gender stereotyping, trait perceptions

---

Introduction

Gender-nonconforming individuals have gender identities and/or expressions that differ from stereotypical gender norms based on one's sex assigned at birth (American Psychological Association, 2018; Turban & Ehrensaft, 2018). Gender-nonconforming children have been shown to have heightened mental health challenges, and poor peer relations is strongly associated with these outcomes (MacMullin et al. 2021; Tankersley et al., 2021; Wittlin et al., 2023). Given this background, there is a need to identify interventions that can improve children's attitudes toward gender nonconformity. Middle childhood in particular may be a key stage to target such interventions toward because past research has found that gender typicality is related to more positive experiences (i.e., lower levels of exclusion and victimization) at this developmental stage (Zosuls et al., 2016).

Previous research that attempted to limit children's gender-biased attitudes is sparse. In an intergroup intervention, children were passively shown vignettes of hypothetical same-aged peers who had both gender-conforming preferences (e.g., girl target liked to jump rope) *and* gender-nonconforming preferences (e.g., boy target liked to do ballet) as well as positive attributes unrelated to gender (e.g., target boy and girl did well in school; Kwan et al., 2020). Exposure to the children in the vignettes was successful at reducing 8- to 9-year-olds' bias against gender-nonconforming peers in a Hong Kong sample (Kwan et al., 2020); however, this same intergroup intervention was not effective in a Canadian sample (MacMullin et al., 2020). An alternate intervention tested by Lamb et al. (2009) used an active, skill-based intervention approach among 5- to 10-year-old children. The intervention focused on teaching children how to confront sexist remarks. Children exposed to the intervention showed evidence of lower levels of gender stereotyping (although some evidence showed that this only occurred in girls) and were more likely to challenge sexist remarks (Lamb et al., 2009). In addition, Bigler and Liben (1992) found that in a social cognitive intervention in which 5- to 10-year-olds were taught multiple classification with social stimuli (i.e., classifying something across more than one dimension [e.g., based on gender and occupation]), they displayed less gender stereotypic views and remembered more gender nonstereotypic information in stories. Overall, findings from several studies indicate that children's gender-related attitudes are malleable and can change in response to intervention techniques; however, successful intervention techniques to limit children's gender-biased attitudes, especially interventions focused on attitudes toward gender-nonconforming peers, are sparse.

Thus, there was mixed effectiveness of a passive intervention approach across studies in Hong Kong and Canada (Kwan et al., 2020; MacMullin et al., 2020), whereas active intervention approaches employed in other research appeared to be effective more consistently (Bigler & Liben, 1992; Lamb et al., 2009). Yet, a disadvantage of active interventions is that they are challenging to implement on a wide scale because they are relatively time-intensive. Developmentally appropriate interventions that are active and skill-based but are less intensive to deliver and can be implemented on a wider scale may be more optimal.

Empathy interventions are good candidates for meeting these criteria. A meta-analysis showed that inducing empathy/perspective-taking can be an effective strategy to improve intergroup relations among children and adolescents (Beelman & Heinemann, 2014). Such interventions are thought to improve intergroup attitudes because empathy for outgroup members is hypothesized to lead individuals to focus more on outgroup members' personal welfare than their group identity (e.g., Batson et al., 1997; Sierksma et al., 2015). For example, Israeli Jewish and Israeli Palestinian children in 3rd and 4th grades showed improved intergroup attitudes (e.g., less discriminatory tendencies, less stereotyping) following an intervention that included training in empathy and perspective-taking (Berger et al., 2016). In another study, when 8- to 13-year-olds were asked how an outgroup member who was in need of help felt (relative to a control group who were not asked this question), children were more likely to report having the intention of helping the outgroup member (Sierksma et al., 2015). Of note, these results held independent of the child's level of trait empathy, as measured by social perspective taking ability. Taken together, past findings indicate that empathy-based intervention approaches appear to be effective among children. Also, children's empathetic concern and perspective-taking skills relate positively to their inclusive attitudes (e.g., Miklikowska, 2018; Nesdale et al., 2005; Taylor et al., 2020). For example, one study found that children's liking for a different ethnic outgroup was positively related to their level of empathy (Nesdale et al, 2005).

Among adult samples, perspective-taking interventions have been shown to improve attitudes about gender diversity. In one randomized control intervention, canvassers went door-to-door in South Florida and spoke with voters (Broockman & Kalla, 2016). In the intervention condition, canvassers asked voters to think about a time when they had been "judged negatively for being different" (p. 221) and consider how this experience might give them an understanding of what transgender people experience. This perspective-taking intervention was successful at improving acceptance of trans people, which held three months after the intervention.

The Present Study

The present study tested an empathy intervention aimed at improving 8- to 11-year-olds' attitudes toward hypothetical gender-nonconforming peers and reducing their gender stereotyping. In this age group, gender typicality is associated with lower levels of both victimization and exclusion (Zosuls et al., 2016). Further, a past intervention technique used in Hong Kong showed it was possible to reduce children's negative attitudes toward gender-nonconforming peers (Kwan et al., 2020), although findings did not hold in a Canadian context (MacMullin et al., 2020), which warrants an examination of different intervention techniques. Empathy-based interventions in particular have been effective at improving intergroup attitudes in past literature (Berger et al., 2016; Sierksma et al., 2015). At this age, children's empathy-related skills are developing and have room to improve (e.g., Lam et al., 2012; Wong et al., 2021). Of note, Wong et al. (2021) found

that children aged 5- and 6-years-old showed levels of empathic accuracy that were comparable to those observed in the adult literature. The current empathy intervention aimed to activate perspective-taking (i.e., imagining being in the target's position) and empathetic understanding/accuracy (i.e., understanding how the target feels; Gleason et al., 2009; Sierksma et al., 2015), which have been shown to be effective at improving intergroup attitudes in previous developmental research (Berger et al., 2016; Sierksma et al., 2015) and in research with adults in the domain of gender diversity (Broockman & Kalla, 2016).

Our preregistered hypothesis was that the intervention (vs. control) condition would improve children's appraisals of hypothetical gender-nonconforming (vs. gender-conforming) peers and reduce children's gender stereotyping, respectively. Importantly, we considered whether any effect of the intervention (vs. control) condition was moderated by child or target peer gender. We did so because past research suggested that children, especially boys, react less positively to gender-nonconforming boy (vs. girl) peers (e.g., Kwan et al., 2020; Nabbijohn et al., 2020; Qian et al., 2021; Skočajić et al., 2020; Wang et al., 2022). Also, boys (vs. girls) could be more resistant to changing their gender stereotyping in response to the intervention, similar to what was found in a prior intervention study (Lamb et al., 2009).

In an exploratory way, we considered possible moderating effects of children's levels of trait empathy on their appraisals of gender-nonconforming (compared with conforming) peers. Children's trait empathy and perspective-taking have been positively associated with their inclusive attitudes (e.g., Miklikowska, 2018; Nesdale et al., 2005). At the same time, one past study found that trait empathy did not moderate the effectiveness of an empathy-based intervention among 8- to 13-year-olds (Sierksma et al., 2015). In the present study, we evaluated whether higher trait empathy would facilitate the intervention's effectiveness or, alternatively, leave less room for the intervention to improve empathetic skills and appraisals of gender-nonconforming (compared with conforming) peers.

Method

Ethics statement

This study was approved by the last author's institutional research ethics board.

Participants

The project was pre-registered with Open Science Framework (OSF; MacMullin & VanderLaan, 2024; https://osf.io/m7js3). To recruit participants, we used Facebook and Instagram advertisements, shared study information through public/community groups, and used a database of local families interested in participating in research maintained in the last author's academic department. We targeted our recruitment toward Ontario, Canada, although participants did not have to be from a certain location.

The study was conducted between January to July of 2021. *A priori*, we aimed to collect data from approximately 180 8- to 11-year-old children (see power analysis details below) with relatively equal numbers of boy and girl participants. Each participant in the study was accompanied by a parent or caregiver who was at least 18 years of age. We pilot tested the study with five participants. The final sample of test participants without pilot test participants included 186 children and their parents/caregivers. There were five participants who were excluded from the final sample due to Internet connection problems, concern about the topic being studied, or lack of assent. In the final sample, all participants were from Canada aside from two participants from the United States. Demographic information about the participants in the current sample are included in Table 1.

Procedure

Parents/caregivers and their children were invited to take part in an online study. Participants were asked to take part in a video call using either Zoom or Skype based on their preference. Upon beginning the video call, the parent met with an experimenter. Experimenters for all participants (including pilot and non-pilot test participants) included five gender-conforming (in terms of gender role expression) women who were undergraduate students of varying races. At this point, the child was asked to leave the room and/or not look at the computer. The parent was asked to wear headphones while the experimenter explained the study procedure and obtained informed consent. Once the parent provided digital consent, they were asked to complete a questionnaire comprised of several measures, including demographic questions about the child and themselves (e.g., the child's age, ethnicity, family composition, socioeconomic status, parents' size of area of residence, education backgrounds, and household income).

Once the parent completed this step, the child was asked to take part in their section of the study. At this point, the investigator shared their screen and performed some initial audio testing to ensure the child could hear the materials presented. Then, the parent was asked to leave the room and/or not look at the computer. The child was asked to wear headphones.

Random assignment was used to designate children to the neutral control condition or empathy intervention condition. Experimenters typically ran a child through the condition opposite to the one that the last child of the same gender went through, with adjustments made as needed to ensure the experimenters ran a relatively even number of participants per gender and condition. Participants were presented with illustrations and associated audio clips that were shared on the investigator's screen. The skin tones of the targets presented were matched as closely as possible to the participant's skin tone, with four options ranging from very light to very dark skin tones, as done in similar previous research (MacMullin et al., 2020; Nabbijohn et al., 2020).

Following a similar design to previous research (e.g., Sierksma et al., 2015), in both the control and intervention conditions, children were shown the same three sets of standardized audiovisual vignettes (see vignettes in Appendix C). Participants first viewed an illustration that introduced the two main child characters in the vignettes (Olivia and Ethan). Next, participants saw three short

Table 1. Descriptive Statistics for Demographic Variables based on Condition and Gender

Variables	Control		Intervention	
	Boys	Girls	Boys	Girls
Age (in years)				
n	44	45	48	49
M (SD)	9.64 (1.06)	9.60 (.99)	9.73 (1.14)	9.55 (1.12)
Parent[a] gender, n (%)				
Man	5 (11.4)	6 (13.3)	8 (16.7)	10 (20.4)
Woman	39 (88.6)	39 (86.7)	40 (83.3)	39 (79.6)
Annual income, n (%)				
< $24,000	0 (0)	1 (2.3)	0 (0)	0 (0)
$24,000 – $50,000	1 (2.3)	3 (7.0)	3 (6.5)	2 (4.3)
$50,000 – $80,000	5 (11.4)	2 (4.7)	6 (13.0)	8 (17.0)
$80,000 – $125,000	16 (36.4)	9 (20.9)	12 (26.1)	8 (17.0)
More than $125,000	22 (50.0)	28 (65.1)	25 (54.3)	29 (61.7)
Parent 1's[a,b] Education, n (%)				
Less than high school	0 (0)	0 (0)	0 (0)	0 (0)
Some high school	0 (0)	0 (0)	0 (0)	0 (0)
High school diploma or equivalent	2 (4.5)	1 (2.2)	1 (2.1)	2 (4.1)
College or Trade	12 (27.3)	7 (15.6)	6 (12.5)	7 (14.3)
University, Bachelors	20 (45.5)	23 (51.1)	24 (50.0)	21 (42.9)
University, Masters	5 (11.4)	6 (13.3)	8 (16.7)	7 (14.3)
University, post-graduate	5 (11.4)	6 (13.3)	5 (10.4)	9 (18.4)
Other professional degree	0 (0)	2 (4.4)	4 (8.3)	3 (6.1)
Relationship Partner Education, n (%)				
Less than high school	0 (0)	0 (0)	0 (0)	0 (0)
Some high school	0 (0)	1 (2.4)	0 (0)	0 (0)
High school diploma	6 (14.6)	3 (7.3)	3 (6.5)	4 (9.3)
College or Trade	16 (39.0)	12 (29.3)	15 (32.6)	15 (34.9)
University, Bachelors	12 (29.3)	12 (29.3)	16 (34.8)	14 (32.6)
University, Masters	4 (9.8)	7 (17.1)	9 (19.6)	8 (18.6)
University, post-graduate	2 (4.9)	5 (12.2)	2 (4.3)	2 (4.7)
Other professional degree	1 (2.4)	1 (2.4)	1 (2.2)	0 (0)
Parent's[a] marital status, n (%)				
Married	36 (81.8)	37 (82.2)	44 (91.7)	41 (83.7)
Separated/divorced	6 (13.6)	2 (4.4)	2 (4.2)	5 (10.2)
Single	0 (0)	1 (2.2)	0 (0)	1 (2.0)
Widowed	1 (2.3)	0 (0)	0 (0)	0 (0)
Common law	1 (2.3)	5 (11.1)	2 (4.2)	2 (4.1)
Area type, n (%)				
Population of less than 1,000	1 (2.3)	0 (0)	1 (2.1)	1 (2.0)
Population between 1,000-9,999	4 (9.1)	5 (11.1)	4 (8.3)	4 (8.2)
Population between 10,000-29,999	5 (11.4)	4 (8.9)	4 (8.3)	1 (2.0)
Population between 30,000-99,999	3 (6.8)	1 (2.2)	3 (6.3)	1 (2.0)
Population between 100,000-499,999	11 (25.0)	16 (35.6)	15 (31.3)	19 (38.8)
Population between 500,000-999,999	8 (18.2)	11 (24.4)	8 (16.7)	11 (22.4)
Population of 1,000,000 or more	12 (27.3)	8 (17.8)	13 (27.1)	12 (24.5)
Ethnicity, n (%)				
North American Indigenous origins	1 (2.3)	1 (2.2)	0 (0)	1 (2.0)
Other North American origins	6 (13.6)	10 (22.2)	8 (16.7)	4 (8.2)
European origins	21 (47.7)	21 (46.7)	22 (45.8)	25 (51.0)
Caribbean origins	0 (0)	0 (0)	1 (2.1)	0 (0)
Latin, Central, and South American origins	0 (0)	0 (0)	0 (0)	1 (2.0)
African origins	0 (0)	0 (0)	0 (0)	0 (0)
Asian origins	7 (15.9)	4 (8.9)	5 (10.4)	9 (18.4)
Oceania origins	0 (0)	0 (0)	0 (0)	0 (0)
Multiple origins	8 (18.2)	9 (20.0)	12 (25.0)	8 (16.3)
Another origin	1 (2.3)	0 (0)	0 (0)	1 (2.0)
Religion, n (%)				
Roman Catholic	10 (22.7)	16 (35.6)	14 (29.2)	11 (22.4)
Protestant	3 (6.8)	5 (11.1)	2 (4.2)	5 (10.2)
Christian	8 (18.2)	9 (20.0)	12 (25.0)	11 (22.4)
Muslim	1 (2.3)	0 (0)	1 (2.1)	4 (8.2)
Jewish	1 (2.3)	0 (0)	0 (0)	1 (2.0)
Buddhist	0 (0)	1 (2.2)	0 (0)	2 (4.1)
Hindu	2 (4.5)	2 (4.4)	2 (4.2)	0 (0)
No religion	14 (31.8)	12 (26.7)	16 (33.3)	9 (18.4)
Multiple religions	4 (9.1)	0 (0)	1 (2.1)	5 (10.2)
Another religion	1 (2.3)	0 (0)	0 (0)	1 (2.0)

Note. M = mean, SD = standard deviation
[a] *Parents here include genetic parents and an aunt.*
[b] *"Parent 1" refers to the participating parent.*

stories depicting these characters, a boy and a girl, who received negative reactions from their (1) peers, (2) teachers, and (3) parents because they had gender-nonconforming preferences in the domains of toy, activity, future career, school subject, dress-up, and chore preferences. The order of the three vignettes (i.e., peer, teacher, and parent vignettes) were randomized across participants. The preferences of the gender-nonconforming children in the vignettes were selected based on previous developmental studies where items relevant to children were rated based on how gender-stereotypical they were considered to be (Blakemore & Centers, 2005; Liben & Bigler, 2002). The aim was for children to have an overall, generalized impression of Olivia and Ethan as having gender-nonconforming preferences. Olivia and Ethan were not depicted as transgender and instead were depicted as gender-nonconforming. All of the illustrations were accompanied by audio narratives, which were narrated by a woman speaking in a neutral tone. Each set of two images (one of Olivia and one of Ethan presented next to each other at the same time) and associated narratives were presented one at a time to participants (e.g., "Olivia's favourite toy is her Spiderman and Ethan's favourite toy is his tea set. Olivia's favourite activity is hockey and Ethan's favourite activity is making jewellery. At school, Olivia and Ethan get teased and excluded by their peers. Olivia's peers tease her because they think Olivia likes toys and activities that are for boys, and Ethan's peers tease him because they think he likes toys and activities that are for girls.")

After being presented with a new piece of information in the vignettes, participants were asked a series of attention check questions (e.g., "What was Olivia's favourite toy?") to ensure they were paying attention to the information presented. In the event that the child was not paying attention, the corresponding illustration and audio clip was replayed. Children were given three chances to get each attention check question correct. If a participant got an attention check question incorrect on their third attempt, it was considered a failed attempt. Participants were excluded from the study if they accumulated three failed attempts on attention check questions about the introduction illustration and the three short stories. No participants were excluded for failing attention check questions.

To manipulate empathy, we used instructions that were given to participants before watching each vignette as well as a question-and-answer period that followed each of the short stories. Prior to watching each short vignette, in the empathy intervention condition, participants were instructed to focus on how the children in the story felt (i.e., "While you are listening, try to think about how the children in the story feel. Pay attention to how the children in the story might feel."). In the control condition, participants were instructed to focus on all the information presented to them (i.e., "While you are

listening, try to observe the children in the story. Pay attention to everything presented to you."). These approaches were similar to past research (e.g., Batson et al., 1997).

Next, in the empathy condition, children were asked questions that led them to focus on the emotions of the targets. Specifically, children in the empathy condition were first asked what Olivia's and Ethan's peers, teacher, and parents told them about their preferences. The corresponding vignette was replayed if children got this question incorrect. In total, children were given up to three chances to answer this question correctly. After this first question, participants were asked how they thought Olivia and Ethan felt at the end of the story on a 5-point scale ranging from 1 = *very sad* to 5 = *very happy* alongside emojis with corresponding emotional expressions. Lastly, participants in the empathy condition were asked to reflect on how they would feel if they experienced a similar event to Olivia and Ethan (e.g., their peers teased and excluded them because of the toys and activities they like), with answer options on a 5-point scale ranging from 1 = *very sad* to 5 = *very happy* alongside emojis with corresponding emotional expressions. Analyses of responses to the final two items appear below in the section labelled Manipulation Check and indicated the manipulation was effective.

In the control condition, children were asked questions that led them to focus on neutral information in the vignettes. Specifically, children were asked who was introduced first in the story, with two answer options of Olivia or Ethan. The corresponding vignette was replayed if children answered this question incorrectly. In total, children were given up to three chances to answer this question correctly. Following, children in the control condition were asked how long the story seemed, with answer options on a 5-point scale ranging from 1= *very short* to 5 = *very long*. Answer options were presented alongside circles of increasing sizes. Lastly, participants were asked when they first heard Olivia and Ethan's names, with five answer options including, for example, near the beginning of the story or near the end of the story.

Following either condition, children completed several dependent measures, including a measure of their ratings of four hypothetical peers who differed in their gender expression and identity followed by a measure of gender stereotyping (note: these peers were different than Olivia and Ethan). At the end of the child's portion of the study, all participants were reminded that it is not fair to judge anyone, exclude them, or make negative comments about them because of what they like to do, no matter their gender. Following, parents were invited back and filled out a debriefing form.

Measures

Trait Perception Ratings of Hypothetical Peers

We assessed participant's attitudes toward gender-nonvconforming peers via a trait perception measured adapted from Martin et al. (2017). Participants were shown four new hypothetical target peers, including a gender-conforming boy, gender-nonconforming boy, gender-conforming girl, and gender-nonconforming girl, which were presented in a randomized order.

These targets have been used in previous research (Kwan et al., 2020; MacMullin et al., 2020; Nabbijohn et al., 2020). The four targets varied in their preferences with respect to toys (i.e., Barbie dolls or miniature race cars, trucks, and planes), activities (i.e., house in a toy kitchen or tackle football), dress-up (i.e., dress-up like their mom by wearing a dress and jewelry or dress-up like their Dad by wearing a suit and tie), and playmates (i.e., best friends who are girls or boys). An audio recorded script, similar to what was used in past research (Kwan et al., 2020; MacMullin et al., 2020; Nabbijohn et al., 2020), describing each target was played alongside the images.

Participants were asked to rate the target's group ("girls like this" or "boys like this"; Martin et al., 2017; Peragine et al., 2021) on a variety of traits. The traits used were selected given that previous research evidence shows that children display an in-group bias favoring their own gender on similar traits (e.g., Martin et al., 2017; Peragine et al., 2021; Powlishta, 1995). Participants were presented with response scales that had contrasting traits on either end, including *very dumb* to *very smart*; *very mean* to *very nice*; *often lie* to *often tell the truth*; *very annoying* to *very friendly*; *very boring* to *very fun*. Participants were asked to respond on a scale of 1–5, with higher scores indicating more positive trait perceptions. As an example, for the gender-conforming boy, participants were asked if "boys like this" were *very annoying*, *a little annoying*, *neither annoying nor friendly*, *a little friendly*, or *very friendly*. The images, script, and rating scales for the trait perception measure are included in Appendix A.

With five items per target child, there were 20 items on the scale overall. We used a mean score for each of the target children on their trait ratings. Cronbach's alphas scores for the ratings of the gender-conforming boy, gender-nonconforming boy, gender-conforming girl, and gender-nonconforming girl were as follows: .78, .74, .80, and .75.

Gender Stereotyping

We used the activities subscale of the Children's Occupations, Activities, and Trait—Attitude Measure (Liben & Bigler, 2002) to assess children's level of gender stereotyping. Specifically, children were presented with 15 questions (shortened scale was used from Fast & Olson, 2017) about whether *only boys*, *only girls*, or *everyone* should do a variety of activities. Each question had an associated illustration that was presented to the participant. Five of the activities were stereotypically associated with girls (e.g., babysit, bake cookies), five of the activities were stereotypically associated with boys (e.g., fly a model plane, play video games), and five were neutral (e.g., play cards, do crossword puzzles).

We scored the measure to assess the proportion of stereotypical answers that children provided (Liben & Bigler, 2002). Specifically, for masculine gender stereotyping, the outcome measure was the proportion of times that the participant responded *only boys* to questions asking about who could do activities stereotypically associated with boys. For feminine gender stereotyping, the outcome measure was the proportion of times (out of 5) that the participant responded *only girls* to questions asking about who could do activities stereotypically associated with girls. Scores ranged

from 0–1, with higher scores indicating higher levels of gender stereotyping (Liben & Bigler, 2002). The Cronbach's alphas for the feminine and masculine stereotyping measures were .76 and .83, respectively.

Trait Empathy

Trait empathy was measured using the Griffith Empathy Measure (Dadds et al., 2008). The Griffith Empathy Measure is a parent-report scale that includes 23 items and is used to gauge children's level of trait empathy (e.g., "My child gets sad when watching sad movies or TV"). Each question is asked on a scale of 1 to 9, where 1 = *strongly disagree* and 9 = *strongly agree*. A mean score was calculated such that a higher score on this scale represented higher levels of trait empathy in the child. The Cronbach's alpha for the trait empathy measure was .86.

Statistical Analysis

The study sample size was determined *a priori* using G*power (Faul et al., 2009). We used repeated-measures analysis of variance (ANOVA) tests with between- and within-subjects interactions. We powered the study to be sensitive to detect small-to-medium effect sizes (f = .20) given that the effects observed in Beelman and Heinemann's (2014) meta-analysis of interventions that have improved intergroup attitudes in children and adolescents were in this range (i.e., d = 0.30 overall, d = 0.44 for empathy/perspective taking interventions). Power was set at a conventional level of .80 (Cohen, 2013) and critical *alpha* was set at .025 (see the last paragraph of this section for the rationale behind selecting this alpha level).

The required sample size was 168 participants for the gender stereotyping measure. The trait perception measure required a smaller sample size of 120 for a .80 power level with an alpha of .025. Given the possibility of some data loss, the *a priori* target sample size was rounded up to approximately 180. With our final sample of 186 participants and observed correlations between repeated measures (see Table 3), we had .80 power to detect a small effect size of f = .07 for the gender stereotyping measure and f = .09–.13 for the trait perception measure (Cohen, 1988). For all models described below, the child's gender was coded as 0 = boy and 1 = girl, and the condition was coded as 0 = control, 1 = intervention.

As stated in our preregistration (MacMullin & VanderLaan, 2024; https://osf.io/m7js3), we ran two mixed-model ANOVAs to test our first two hypotheses. To test whether the intervention (vs. control) condition improved children's appraisals of hypothetical gender-nonconforming (vs. gender-conforming) peers and whether any condition effects were moderated by child or target peer gender, we ran a mixed-model ANOVA with trait perception ratings of target peers set as the dependent measure. The independent variables included two within-subjects factors, which were the target gender (boy vs. girl) and the target gender expression (gender-conforming vs. gender-nonconforming). Also, we included two between-subjects factors, which were the child's gender and the condition. To test whether the intervention

Table 2. Descriptive Statistics for Gender Stereotyping and Trait Perception by Condition and Child Gender

	Control		Intervention	
	Boys	Girls	Boys	Girls
	(n = 44)	(n = 45)	(n = 48)	(n = 49)
Gender Stereotyping[a] M (SD)				
Feminine	.21 (.28)	.08 (.18)	.22 (.29)	.23 (.31)
Masculine	.31 (.35)	.11 (.25)	.26 (.34)	.24 (.31)
Trait Perception[b] M (SD)				
Gender-conforming boy	3.94 (.68)	3.64 (.83)	3.71 (.68)	3.67 (.77)
Gender-nonconforming boy	3.89 (.72)	4.07 (.74)	3.71 (.68)	3.93 (.74)
Gender-conforming girl	4.00 (.70)	3.68 (.89)	3.68 (.77)	4.07 (.72)
Gender-nonconforming girl	3.90 (.71)	3.98 (.73)	3.69 (.58)	4.01 (.71)

[a] *Measured by the Children's Occupations, Activities, and Trait—Attitude Measure (Liben & Bigler, 2002) shortened version (Fast & Olson, 2017). The feminine and masculine stereotyping scores capture the proportion of stereotypical responses children gave to questions about activities stereotyped for girls and activities stereotyped for boys, where higher scores indicate higher levels of gender stereotyping.*
[b] *Measured by a trait perception measure adapted from Martin et al., 2017. Higher scores indicate more positive appraisals of the target children.*

(vs. control) condition reduced children's gender stereotyping and whether any condition effects were moderated by child or target peer gender, we ran a mixed-model ANOVA with gender stereotyping as the dependent measure. The independent variables included a within-subjects factor of the stereotype category (i.e., feminine vs. masculine). Also, we included two between-subjects factors, which were the child's gender and the condition. Post-hoc ANOVAs, *t*-tests, and Pearson's correlations were run to further examine highest-order significant effects where they were found.

In an exploratory way, we examined whether the child's trait empathy levels moderated any effects on the children's trait perception ratings. For completeness, we also examined whether the child's trait empathy levels moderated any effects on the children's level of gender stereotyping, which is included in Appendix B in Table A1 and is not detailed further below. These analyses were not included in our preregistration. We ran mixed-model ANOVAs with trait perception ratings set as the dependent measure. We were only interested in interpreting effects related to trait empathy. The models were the same as those described above, but we included trait empathy as a continuous factor and modeled its interaction with the other factors in the model.

For all mixed-model ANOVAs described above, we applied a Bonferroni correction to our *p*-values. A *p*-value was considered significant if it was less than .025—i.e., a conventional .05 divided by two given two outcome variables in our preregistered analyses (i.e., gender stereotyping, trait perception ratings). We considered a *p*-value to be significant it was less than .025 (i.e., .05 divided by two) for our post-hoc tests. This was done because, for post-hoc tests, we split the data by the number of levels of the independent variable and this number was always two. When reporting our results, redundant effects/comparisons are not repeated.

Results

Descriptive statistics for the trait perception and gender stereotyping measures are presented in Table 2, which includes trait perception ratings of gender-conforming and gender-

nonconforming targets by child gender and condition. Table 3 presents zero-order correlations between measures in the study divided by child gender. For analyses demonstrating that the intervention manipulation was effective at eliciting empathy and perspective-taking using one-sample t-tests, refer to the Results section below.

Manipulation Check

In the intervention condition, children were asked: "How did Olivia and Ethan feel at the end of the story?", with five answer options ranging from *very sad* to *very happy* at the end of each of the three short stories. As a means of ensuring that children in the intervention condition were attuned to how Olivia and Ethan likely felt at the end of each story, we ran three one-sample t-tests comparing responses of children in the intervention condition to the scale midpoint (i.e., 3). We expected that children's responses would be significantly below the scale midpoint (i.e., responding with *very sad* or *a little sad*). The mean (*SD*) for the question applied to Olivia and Ethan's peers, teacher, and parents, respectively, were 1.33 (.55), 1.38 (.60), and 1.42 (.56). For each question, children's responses were significantly below the scale midpoint, all ts < −26.44 and all ps < .001. Overall, > 95% of participants responded *a little sad* or *very sad* for all questions.

Also, in the intervention condition, children were asked how they would feel if they experienced a similar event to Olivia and Ethan (e.g., their peers teased and excluded them because of the toys and activities they like), with answer options on a 5-point scale ranging from 1 = *very sad* to 5 = *very happy* alongside emojis with corresponding emotional expressions. As a means of ensuring that children in the intervention condition were aware of how they would feel if they experienced a similar event to Olivia and Ethan, we ran three one-sample t-tests comparing responses of children in the intervention condition to the scale midpoint (i.e., 3). This analysis was not included in our preregistration. We expected that children's responses would be significantly below the scale midpoint (i.e., responding with *very sad* or *a little sad*). The mean (*SD*) for the question applied to Olivia and Ethan's peers, teacher, and parents, respectively, were 1.38 (.57), 1.45 (.68), and 1.58 (.67). For each question, children's responses were significantly below the

scale midpoint, all ts < −20.78 and all ps < .001. Overall, > 91% of participants responded *a little sad* or *very sad* for all questions.

Demographic Analysis

We used chi-square tests to explore whether demographic variables (see Table 1) differed by condition and/or child gender and should therefore be controlled in our analyses. We examined the relations between condition and child gender with ethnicity (0 = European origins, 1 = all other origins, including multiple origins or "another" origin), area (0 = population below 500,000, 1 = population over or equal to 500,000), income (0 = less than $125,000, 1 = more than $125,000), and religion (0 = Roman Catholic, Protestant, Christian, 1 = all other religions, including no religion, multiple religions, or "another" religion). There were no demographic variables that differed by condition, all Pearson's chi-square values < .05, all ps > .833. Also, there were no demographic variables that differed by child gender, all Pearson's chi-square values < 1.72, all ps > .190.

Preregistered Analyses Testing Intervention Effects: Role of Condition on Trait Perception

The results of the mixed-model ANOVA for the trait perception and gender stereotyping measures are displayed in Table 4 and Table 5, respectively. As shown in Figure 1, no significant effects were found for the intervention in a way that conformed to our hypotheses/predictions (i.e., no significant effects involving Target Gender Expression × Condition for trait perception ratings and no significant effects involving condition on gender stereotyping). Any additional findings are detailed in Appendix B and descriptive statistics are displayed in Table 2.

Exploratory Analyses
Role of Trait Empathy on Trait Perception

The results of the mixed-model ANOVA for the trait perception measure including trait empathy as a continuous predicting factor and modelling its interaction with other factors reported in the prior section are displayed in Table 6. As shown in Table 6, there was a significant Child Gender × Empathy interaction, $F(1, 178) = 9.25$, $p = .003$, $\eta_p^2 = .05$. As displayed in Figure 2, among girls, having a higher level of trait empathy was associated with more positive ratings toward target children on the trait perception measure, $r = .35$, $p < .001$. The relationship between the child's level of trait empathy and trait perception scores was nonsignificant in boys, $r = .02$, $p = .860$.

Discussion

This study investigated relations among an empathy manipulation, trait empathy, gender stereotyping, and trait perceptions of target children to provide insight into how children's

Table 3. Zero-Order Correlations for Outcome Variables

	1	2	3	4	5	6
1. Feminine Stereotyping		.89**	0.18	−.24*	0.14	−0.14
2. Masculine Stereotyping	.86**		0.19	−.29**	0.17	−0.19
3. Trait Perception – Gender-Conforming Boy	−0.07	<.01		.25*	.78**	.45**
4. Trait Perception – Gender-Nonconforming Boy	−0.14	−.22*	.43**		.35**	.62**
5. Trait Perception – Gender-Conforming Girl	.26*	.27**	.65**	.32**		.44**
6. Trait Perception – Gender-Nonconforming Girl	−0.03	−0.02	.58**	.71**	.50**	

* p-value < .05 (two-tailed)
** p-value < .01 (two-tailed)

Table 4. Mixed-Model ANOVA for Trait Perception Ratings for Gender-Conforming Boy and Girl and Gender-Nonconforming Boy and Girl

	F-value	*p*	Partial Eta-Squared
Target Gender	3.98	0.048	0.02
Target Gender × Child Gender	2.98	0.086	0.02
Target Gender × Condition	3.34	0.069	0.02
Target Gender × Child Gender × Condition	7.91	0.005	0.04
Target gender expression	3.80	0.053	0.02
Target Gender Expression × Child Gender	7.21	0.008	0.04
Target Gender Expression × Condition	0.89	0.348	0.01
Target Gender Expression × Child Gender × Condition	3.10	0.080	0.02
Target Gender × Target Gender Expression	3.86	0.051	0.02
Target Gender × Target Gender Expression × Child Gender	2.53	0.113	0.01
Target Gender × Target Gender Expression × Condition	0.27	0.608	<.01
Target Gender × Target Gender Expression × Child Gender × Condition	1.05	0.307	0.01
Child gender	0.62	0.432	<.01
Condition	0.78	0.379	<.01
Child Gender × Condition	3.39	0.067	0.02

Table 5. Mixed-Model ANOVA for Feminine and Masculine Gender Stereotyping

	F-value	*p*	Partial Eta-Squared
Stereotype category	15.36	<.001	0.08
Stereotype Category × Child Gender	5.61	0.019	0.03
Stereotype Category × Condition	2.30	0.131	0.01
Stereotype Category × Child Gender × Condition	0.48	0.489	<.01
Child gender	4.23	0.041	0.02
Condition	1.86	0.174	0.01
Child Gender × Condition	3.57	0.060	0.02

(a)

(b)

Figure 1. Trait Perception Ratings and Gender Stereotyping Ratings by Condition

Note. Error bars show 95% confidence intervals. Figure 1 (a) shows the null trait perception findings by plotting ratings of gender conforming and gender-nonconforming targets per condition. Figure 1 (b) shows the null gender stereotyping findings by plotting feminine and masculine gender stereotyping per condition. In (a), higher scores represent more positive appraisals of the target children. Figure 1 (b), higher scores represent higher levels of gender stereotyping.

Figure 2. Trait Empathy and Trait Perception Ratings by Child Gender
Note. Boys (light blue line) and girls (dark green line) shown separately. Higher trait perception ratings represent more positive appraisals of the target children. Higher trait empathy scores represent higher levels of trait empathy.

gender-related attitudes can be improved. The empathy-based intervention evaluated here was not effective at reducing children's gender stereotyping or improving their perceptions of gender-nonconforming children. We did find that children's perceptions of the peers' traits were moderated by trait empathy. Among girls, but not boys, trait empathy scores were positively associated with more positive attitudes towards all target children. Together, these findings can help inform future work aimed at improving children's gender-related attitudes.

Notably, the empathy intervention evaluated in the current study was based on content and principles that were established in prior literature—including a meta-analysis (Beelmann & Heinemann, 2014) and individual studies (e.g., Berger et al., 2016; Bigler & Liben, 1992; Kwan et al., 2020; Lamb et al., 2009; MacMullin et al., 2020; Sierksma et al., 2015) focused on children and youth—the sample provided adequate statistical power, and the preregistered analytic plan was centered on evaluating the intervention. It is possible the literature supporting
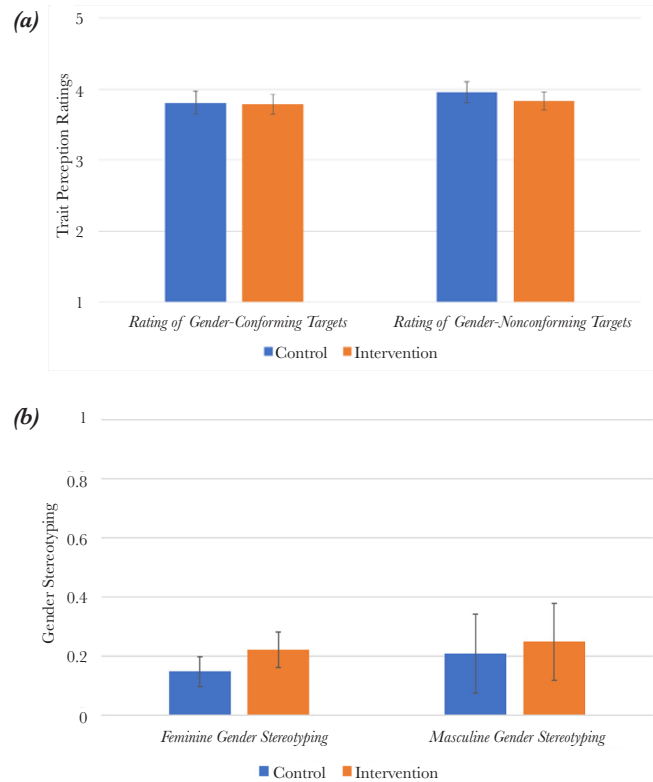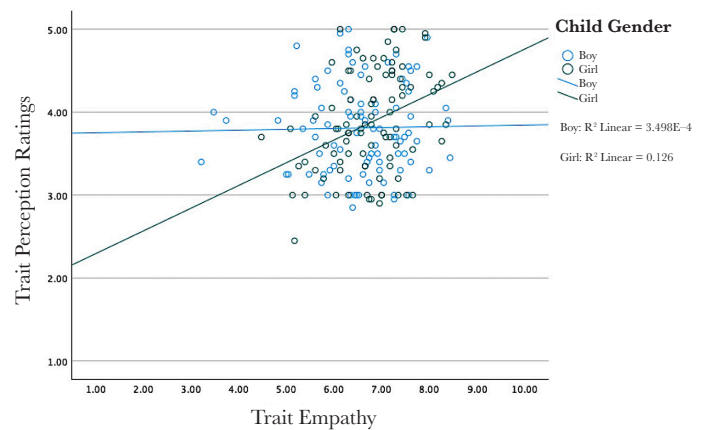
empathy interventions is biased toward positive and significant results due to a publication bias of significant over null findings in the primary literature (see Paluck et al., 2021). If so, the present intervention might have produced null effects because—despite the strengths of the study design—empathy interventions on intergroup attitudes and stereotyping are in fact less effective than the published literature suggested. Another possibility is that empathy interventions aimed at children are not effective in the domain of gender stereotyping and attitudes toward gender-nonconforming peers relative to other domains tested in previous research (e.g., willingness to help an outgroup member). It is not immediately obvious why empathy interventions would be less effective in the domain of gender; however, gender may take precedence over other kinds of social categories for children (e.g., race; Qian et al., 2021) and, therefore, be more resistant to change (but see Bigler & Liben, 1992; Kwan et al., 2020; Lamb et al., 2009).

Relatedly, in certain domains of gender-related attitudes (e.g., peer preference based on gender, holding biased attitudes favoring one's own-gender), children show high degrees of rigidity throughout middle childhood (Halim, 2016). Thus, the present study may not have been effective because children remain especially rigid in their gender-typed thinking in categories relevant to the current study during middle childhood. As noted by Halim (2016), this rigidity in middle childhood in certain domains of gender-related attitudes may be related to "stringent peer enforcement of gender norms" (p. 158). Another possible reason why the intervention was not effective relates to the fact that children who experience high pressure to be gender-conforming might infer that they will experience social costs for affiliating with gender-nonconforming peers (Masters et al., 2021). Thus, in the present study, if children felt high levels of pressure to be gender-conforming, they may have been resistant to changing their attitudes due to concerns about repercussions.

Another possibility is that the intervention tested here was not active enough (see Bigler & Liben, 1992; Broockman & Kalla, 2016; Lamb et al., 2009). For example, unlike some previous studies (e.g., Broockman & Kalla, 2016), there was limited back-and-forth conversation between the child participant and the experimenter. In a similar vein to the procedures of Broockman and Kalla (2016), the intervention may have been improved if children were asked to recall their own experience of being discriminated against in order to arouse a more personal experience. Further, it might have been effective to give children more opportunities to practice engaging in empathy throughout the intervention (e.g., more repetitions of the short stories with follow-up questions). In addition, a future intervention may focus on targeting the belief that gender is a domain in need of policing directly (e.g., through directly challenging those kinds of thoughts). Overall, effective interventions for children to improve their gender-related attitudes that are not overly time

Table 6. Mixed-Model ANOVA for Trait Perception Ratings for Gender-Conforming Boy and Girl and Gender-Nonconforming Boy and Girl with Trait Empathy

	F-value	p	Partial Eta-Squared
Target gender	0.01	0.923	<.01
Target Gender × Condition	1.31	0.253	0.01
Target Gender × Child Gender	2.68	0.103	0.02
Target Gender × Empathy	0.10	0.753	<.01
Target Gender × Condition × Child Gender	0.64	0.426	<.01
Target Gender × Condition × Empathy	0.84	0.361	0.01
Target Gender × Child Gender × Empathy	3.52	0.062	0.02
Target Gender × Condition × Child Gender × Empathy	0.18	0.675	<.01
Target gender e×pression	0.32	0.574	<.01
Target Gender E×pression × Condition	1.10	0.296	0.01
Target Gender E×pression × Child Gender	0.12	0.732	<.01
Target Gender E×pression × Empathy	0.10	0.751	<.01
Target Gender E×pression × Condition × Child Gender	1.44	0.231	0.01
Target Gender E×pression × Condition × Empathy	0.85	0.359	0.01
Target Gender E×pression × Child Gender × Empathy	<.01	0.995	<.01
Target Gender E×pression × Condition × Child Gender × Empathy	0.95	0.331	0.01
Target Gender × Target Gender E×pression	0.01	0.908	<.01
Target Gender × Target Gender E×pression × Condition	1.46	0.228	0.01
Target Gender × Target Gender E×pression × Child Gender	<.01	0.976	<.01
Target Gender × Target Gender E×pression × Empathy	0.02	0.888	<.01
Target Gender × Target Gender E×pression × Condition × Child Gender	4.80	0.030	0.03
Target Gender × Target Gender E×pression × Condition × Empathy	1.78	0.183	0.01
Target Gender × Target Gender E×pression × Child Gender × Empathy	0.03	0.867	<.01
Target Gender × Target Gender E×pression × Condition × Child Gender × Empathy	4.46	0.036	0.02
Condition	1.37	0.243	0.01
Child gender	8.79	0.003	0.05
Empathy	10.44	0.001	0.06
Condition × Child Gender	1.66	0.199	0.01
Condition × Empathy	1.62	0.205	0.01
Child Gender × Empathy	9.25	0.003	0.05
Condition × Child Gender × Empathy	1.05	0.306	0.01

Note. A p-value was considered significant if it was less than .025.

intensive are needed in order to improve children's attitudes.

Similar to prior research (Sierksma et al., 2015), trait empathy did not moderate the intervention's effectiveness; however, we did find an interaction effect between the child's gender and level of trait empathy in predicting their ratings of hypothetical target peers. Specifically, girl participants had less positive ratings of all target children, independent of target gender or gender expression, when they had lower levels of trait empathy. The same relationship was not found in boy participants. Future research should aim to better understand the role that a child's own gender plays in the relationship between their level of empathy and interpersonal perceptions, as some have previously suggested should be explored further (Miklikowska, 2018). Relatedly, future researchers should examine whether (and, if so, why) for girls,

relative to boys, their attitudes toward peers are impacted more by their level of trait empathy. One recent study found that young female, but not male, child participants in kindergarten and first grade engaged in more behaviors indicative of empathy (e.g., looking at the victim, helping the victim) after witnessing a same-sex classmate experience an adverse event (Benenson et al., 2021). Overall, it is possible that empathy is a key factor for predicting behavior and attitudes toward others more so for girls than for boys; however, more research is needed.

Contrary to prior studies (e.g., Kwan et al., 2020; Nabbijohn et al., 2020; Qian et al., 2021; Skočajić et al., 2020; Wang et al., 2022; Zosuls et al., 2016), we did not find that on average children rated the gender-nonconforming (vs. gender-conforming) target children less positively. We may have observed a lack of negative attitudes toward gender-nonconforming target children in the current study because the particular sample of children in our study had highly accepting attitudes. Assuming our lack of evidence of negative attitudes toward gender nonconformity using the trait perception measure is not a case of Type II error despite a reasonable level of statistical power, it is possible that children are biased against gender-nonconforming children in other domains captured in previous research (e.g., friendship preference, moral judgement), but not in terms of their positive-negative trait attributions, as was measured in the current study.

Limitations and Future Directions

A limitation of the present study, as noted above, is that our intervention may not have been active enough. Those designing interventions in the future might consider incorporating opportunities for the experimenter to ask the participants more follow-up questions (e.g., "How do you feel about what happened to Olivia and Ethan?"; "Why do you think Olivia and Ethan felt that way at the end of the story?"). However, it is worth noting that past empathy interventions that were not very active have been effective at improving children's attitudes toward outgroup members (e.g., Sierksma et al., 2015). A further limitation of the present study is that the results do not generalize beyond Western contexts, which is noteworthy given recent research reporting cultural differences in the effectiveness of interventions related to children's gender attitudes (Kwan et al., 2020; MacMullin et al., 2020) as well as the development of children's gender-related peer appraisals (Kwan et al., 2020; Nabbijohn et al., 2020; Wang et al., 2022) and gender stereotyping (Qian et al., 2023). Lastly, we found that among girls, trait empathy scores were positively associated with more positive attitudes toward all target children. Thus, future intervention work may focus on improving empathy as a means of improving children's attitudes toward peers broadly.

Further, it is possible that children in the present study responded in a socially desirable way on our trait perception measure. To mitigate this issue in the future, researchers may use other measures such as sticker distribution tasks or rank-order tasks. Finally, in the present study, we used a parent-report measure of trait empathy in order to keep the length of the study manageable for children. Future researchers may instead use a child-report measure of empathy, such as the Basic Empathy Scale

(Jolliffe & Farrington, 2006).

Conclusion

The present study tested an empathy intervention aimed at reducing children's gender stereotyping and improving their attitudes toward gender-nonconforming children. Contrary to our hypotheses, the intervention was not effective at reducing children's level of gender stereotyping or improving their trait perceptions of gender-nonconforming children. We did find that trait empathy had a positive association with attitudes toward target children in girls (but not in boys). Although this association was not specific to gender-nonconforming targets, it nevertheless suggests the usefulness of paying more attention to various forms and measures of empathy. Given that poor peer relations contribute to negative mental health outcomes for gender-nonconforming children (MacMullin et al. 2021; Tankersley et al., 2021; Wittlin et al., 2023), future work should focus on building interventions that improve children's attitudes toward gender-nonconforming peers.

Data Availability Statement
We will make our data publicly available upon acceptance.

References

American Psychological Association. (2018, September). A glossary: Defining transgender terms. American Psychological Association. https://www.apa.org/monitor/2018/09/ce-corner-glossary

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H., Mitchener, E. C., Bednar, L. L., Klen, T. R., & Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology, 72,* 105–118. https://doi.org/10.1037/0022-3514.72.1.105

Beelmann, A., & Heinemann, K. S. (2014). Preventing prejudice and improving intergroup attitudes: A meta-analysis of child and adolescent training program. *Journal of Applied Developmental Psychology, 34,* 10–24. https://doi.org/10.1016/j.appdev.2013.11.002

Benenson, J. F., Gauthier, E., & Markovits, H. (2021). Girls exhibit greater empathy than boys following a minor accident. *Scientific Reports, 11,* 7965. https://doi.org/10.1038/s41598-021-87214-x

Berger, R., Benatov, J., Abu-Raiya, H., & Tadmor, C. T. (2016). Reducing prejudice and promoting positive intergroup attitudes among elementary-school children in the context of the Israeli-Palestinian conflict. *Journal of School Psychology, 57,* 53–72. https://doi.org/10.1016/j.jsp.2016.04.003

Bigler, R., & Liben, L. (1992). Cognitive mechanisms in children's gender stereotyping: Theoretical and educational implications of a cognitive-based intervention. *Child Development, 63,* 1351–1363. https://doi.org/10.2307/1131561

Blakemore, J. E., & Centers, R. E. (2005). Characteristics of boys' and girls' toys. *Sex Roles, 53,* 619–633. https://doi.org/10.1007/s11199-005-7729-0

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352,* 220–224. https://doi.org/10.1126/science.aad9713

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates.

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

Dadds, M. R., Hunter, K., Hawes, D. J., Frost, A. D. J., Vassallo, S., Bunn, P., Merz, S., & Masry, Y. E. (2008). A measure of cognitive and affective empathy in children using parent ratings. *Child Psychiatry and Human Development, 39,* 111–122. https://doi.org/10.1007/s10578-007-0075-4

Fast, A., & Olson, K. R. (2017). Gender development in preschool transgender children. *Child Development, 89,* 620–637. https://doi.org/10.1111/cdev.12758

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

Gleason, K. A., Jensen-Campbell, L. A., & Ickes, W. (2009). The role of empathic accuracy in adolescents' peer relations and adjustment. *Personality and Social Psychology Bulletin, 35,* 997–1011. https://doi.org/10.1177/0146167209336605

Halim, M. L. D. (2016). Princesses and superheroes: Social-cognitive influences on early gender rigidity. *Child Development Perspective, 10,* 155–160. https://doi.org/10.1111/cdep.12176

Halim, M. L, D., Sakamoto, D. J., Russo, L. N., Echave, K. N., Portillo, M. A., & Tawa, S. (2022). Early gender differences in valuing strength. *Archives of Sexual Behavior, 51,* 2169–2182. https://doi.org/10.1007/s10508-021-02185-4

Jolliffe, D., & Farrington, D. P. (2006). Basic Empathy Scale (BES) [Database record]. PsycTESTS.

Kwan, K. M. W., Shi, S. Y., Nabbijohn, A. N., MacMullin, L. N., VanderLaan, D. P., & Wong, W. I. (2020). Children's appraisals of gender nonconformity: Developmental pattern and intervention. *Child Development, 91,* e780–e789. https://doi.org/10.1111/cdev.13316

Lam, C. B., Solmeyer, A. R., & McHale, S. M. (2012). Sibling relationships and empathy across the transition to adolescence. *Journal of Youth and Adolescence, 41,* 1657–1670. https://doi.org/10.1007/s10964-012-9781-8

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

Liben, L. S., & Bigler, R. S. (2002). The developmental course of gender differentiation: Conceptualizing, measuring, and evaluating constructs and pathways. *Monographs of the Society for Research in Child Development, 67,* 1–147. https://doi.org/10.1111/1540-5834.t01-1-00187

MacMullin, L. N., Bokeloh, L. M., Nabbijohn, A. N., Santarossa, A., van der Miesen, A. I. R., Peragine, D. E., & VanderLaan, D. P. (2021). Psychological well-being of gender-nonconforming children in a community sample: The roles of peers and parents. *Archives of Sexual Behavior, 50,* 823–841. https://doi.org/10.1007/s10508-020-01832-6

MacMullin, L. N., Nabbijohn, A. N., Kwan, K. M. W., Santarossa, A., Peragine, D. E., James, H. J., Wong, W. I., & VanderLaan, D. P. (2020). Testing an intergroup relations intervention strategy to improve children's appraisals of gender-nonconforming peers. *Canadian Journal of Human Sexuality, 29,* 167–181. https://doi.org/10.3138/cjhs.2020-0001

MacMullin, L., & VanderLaan, D. (2024, November 22). Child emotions study. OSF. https://osf.io/m7js3

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

Masters, S. L., Hixson, K., & Hayes, A. R. (2021). Perceptions of gender norm violations among middle school students: An experimental study of the effects of violation type on exclusion expectations. *The Journal of Early Adolescence, 41,* 527–549. https://doi.org/10.1177/0272431620931193

Miklikowska, M. (2018). Empathy trumps prejudice: The longitudinal relation between empathy and anti-immigrant attitudes in adolescence. *Developmental Psychology, 54,* 703–717. https://doi.org/10.1037/dev0000474

Nabbijohn, A. N., MacMullin, L. N., Kwan, K. M. W., Santarossa, A., Peragine, D. E., Wong, W. I., & VanderLaan, D. P. (2020). Children's bias in appraisals of gender-variant peers. *Journal of Experimental Child Psychology, 196,* 104865. https://doi.org/10.1016/j.jecp.2020.104865

Nesdale, D., Griffith, J., Durkin, K., & Maass, A. (2005). Empathy, group norms and children's ethnic attitudes. *Applied Developmental Psychology, 26,* 623–637. https://doi.org/10.1016/j.appdev.2005.08.003

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology, 72,* 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Patterson, M. M. (2012). Self-perceived gender typicality, gender-typed attributes, and gender stereotype endorsement in elementary-school-aged children. *Sex Roles, 67,* 422–434. https://doi.org/10.1007/s11199-012-0184-9

Peragine, D. E., Gill, B., Troisi, V. V., MacMullin, L. N, & VanderLaan, D. P. (2021). Children's intergroup gender bias and self-perceived same- and other-gender similarity. *British Journal of Developmental Psychology, 39,* 125–132. https://doi.org/10.1111/bjdp.12351

Powlishta, K. K. (1995). Gender bias in children's perceptions of personality traits. *Sex Roles, 32,* 17–28. https://doi.org/10.1007/bf01544755

Qian, M., Wang, Y., Wong, W. I., Fu, G., Zuo, B., & VanderLaan, D. (2021). The effects of race, gender, and gender-typed behavior on children's friendship appraisals. *Archives of Sexual Behavior, 50,* 807–820. https://doi.org/10.1007/s10508-020-01825-5

Qian, M., Wong, W. I., Nabbijohn, A. N., Wang, Y., MacMullin, L. N., James, H. J., Fu, G., Zuo, B., & VanderLaan, D. P. (2023). Children's implicit gender–toy association development varies across cultures. Developmental Psychology, 59(12), 2287–2295. https://doi.org/10.1037/dev0001590

Sierksma, J., Thijs, J., & Verkuyten, M. (2015). In-group bias in children's intention to help can be overpowered by inducing empathy. *British Journal of Developmental Psychology, 33,* 45–56. https://doi.org/10.1111/bjdp.1206

Skočajić, M. M., Radosavljević, J. G., Okičić, M. G., Janković, I. O., & Žeželj, I. L. (2020). Boys just don't! Gender stereotyping and sanctioning of counter-stereotypical behavior in preschoolers. *Sex Roles, 82,* 163–172. https://doi.org/10.1007/s11199-019-01051-x

Spivey, L. A., Huebner, D. M., & Diamond, L. M. (2018). Parent responses to childhood gender nonconformity: Effects of parent and child characteristics. *Psychology of Sexual Orientation and Gender Diversity, 5,* 360–370. https://doi.org/10.1037/sgd0000279

Tankersley, A. P., Grafsky, E.L., Dike, J., & Jones, R. T. (2021). Risk and resilience factors for mental health among transgender and gender Nonconforming (TGNC) youth: A systematic review. *Clinical Child and Family Psychology Review, 24,* 183–206. https://doi.org/10.1007/s10567-021-00344-6

Taylor, L. K., O'Driscoll, D., Dautel, J. B., & McKeown, S. (2020). Empathy to action: Child and adolescent out-group attitudes and prosocial behaviors in a setting of intergroup conflict. *Social Development, 29,* 461–477. https://doi.org/10.1111/sode.12421

Turban, J. L., & Ehrensaft, D. (2018). Research review: Gender identity in youth: Treatment paradigms and controversies. *Journal of Child Psychology and Psychiatry, 59,* 1228–1243.

Wang, Y., Qian, M., Nabbijohn, A. N., Wen, F., Fu, G., Zuo, B., & VanderLaan, D. P. (2022). Culture influences the development of children's gender-related peer preferences: Evidence from China and Thailand. *Developmental Science, 25,* e13221. https://doi.org/10.1111/desc.13221

Wittlin, N. M., Kuper, L. E., & Olson, K. R. (2023). Mental health of transgender and gender diverse youth. *Annual Review of Clinical Psychology, 19,* 207–232. https://doi.org/10.1146/annurev-clinpsy-072220-020326

Wong, W. I., Tsui, W., & Siu, T. C. (2021). Empathic accuracy of young boys and girls in ongoing parent-child interactions: Performance and (mis)perception. *Journal of Experimental Child Psychology, 203,* 105042. https://doi.org/10.1016/j.jecp.2020.105042

Zosuls, K. M., Andrews, N. C. Z., Martin, C. L., England, D. E., & Field, R. D. (2016). Developmental changes in the link between gender typicality and peer victimization and exclusion. *Sex Roles, 75,* 243–256. https://doi.org/10.1007/s11199-016-0608-z

Appendix A

Script for Trait Perception Measure

Gender-conforming boy: "Some boys, like this boy, like to play with miniature race cars, trucks, and planes, play tackle football, dress up like their Dad by wearing a suit and tie, and have boys as best friends."

Gender-nonconforming boy: "Some boys, like this boy, like to play with Barbie dolls, play house in a toy kitchen, dress up like their mom by wearing a dress and jewelry, and have girls as best friends."

Gender-conforming girl: "Some girls, like this girl, like to play with Barbie dolls, play house in a toy kitchen, dress up like their mom by wearing a dress and jewelry, and have girls as best friends."

Gender-nonconforming girl: "Some girls, like this girl, like to play with miniature race cars, trucks, and planes, play tackle football, dress up like their Dad by wearing a suit and tie, and have boys as best friends."

Trait Perception Measure Questions

Boys [or girls] like this are:
> *Very dumb*
> *A little dumb*
> *Not dumb or smart*
> *A little smart*
> *Very smart*

Boys [or girls] like this are:
> *Very mean*
> *A little mean*
> *Not mean or nice*
> *A little nice*
> *Very nice*

Boys [or girls] like this:
> *Often lie*
> *Sometimes lie*
> *Don't lie or tell the truth*
> *Sometimes tell the truth*
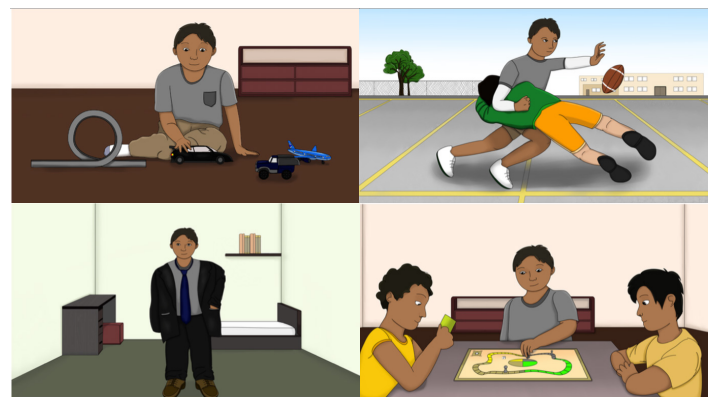> *Often tell the truth*

Boys [or girls] like this are:
> *Very annoying*
> *A little annoying*
> *Not annoying or friendly*
> *A little friendly*
> *Very friendly*

Boys [or girls] like this are:
> *Very boring*
> *A little boring*
> *Not boring or fun*
> *A little fun*
> *Very fun*

Trait Perception Measure Images

Gender-conforming boy (medium dark).

Gender-conforming girl (medium dark).

Gender-nonconforming boy (medium dark).

Gender-nonconforming girl (medium dark).

Appendix B

As shown in Table 4, there was a significant Target Gender × Child Gender × Condition interaction effect, $F(1, 182) = 7.91$, $p = .005$, $\eta_p^2 = .04$ (see Figure A1 a) for the trait perception measure. To follow up on this effect, another mixed-model ANOVA was run separately in boys and in girls to test whether the Target Gender × Condition effect held in the boy and girl groups. Although the Target Gender × Condition effect was not significant in boys, $F(1, 90) = .62$, $p = .434$, $\eta_p^2 = .01$, it was significant in girls, $F(1, 92) = 8.93$, $p = .004$, $\eta_p^2 = .09$. Follow-up independent samples t-tests examined girls' ratings of boy and girl targets, respectively, in the control vs. intervention conditions. Contrary to the interaction effect in the ANOVA, these follow-up tests showed no significant difference in girls' ratings of boy or girl targets across conditions. Girls did not differ significantly in how they rated boy targets in the control ($M = 3.85$, $SD = .69$) vs. intervention condition ($M = 3.80$, $SD = .62$), $t(92) = .40$, $p = .693$, Cohen's $d = .08$. Also, girls did not differ significantly in how they rated girl targets in the control ($M = 3.83$, $SD = .70$) vs. intervention ($M = 4.04$, $SD = .63$) condition, $t(92) = -1.58$, $p = .118$, Cohen's $d = -.33$. More importantly, there was no effect of condition on ratings of gender conforming vs. gender-nonconforming targets, the main comparison of interest.

Initially, we found that the Target Gender × Condition effect was significant in girls, $F(1, 92) = 8.93$, $p = .004$, $\eta_p^2 = .09$. Thus, within girl participants, we split the dataset based on condition and ran paired-samples t-tests to explore how girls rated girl targets vs. boy targets in the control and intervention conditions, respectively. In the control condition, girls did not rate the boy targets ($M = 3.85$, $SD = .69$) or girl targets ($M = 3.83$, $SD = .70$) differently, $t(44) = .43$, $p = .670$, Cohen's $d = .06$. In the intervention condition, girls rated the boy targets ($M = 3.80$, $SD = .62$) less positively than the girl targets ($M = 4.04$, $SD = .63$), $t(48) = -3.74$, $p < .001$, Cohen's $d = -.54$.

A mixed-model ANOVA was run separately in the control and intervention conditions to test whether the Target Gender × Child Gender effect held in the control and intervention conditions. Although the Target Gender × Child Gender effect was not significant in the control condition, $F(1, 87) = .60$, $p = .439$,

$\eta_p^2 = .01$, it was significant in the intervention condition, $F(1, 95) = 10.16$, $p = .002$, $\eta_p^2 = .10$. Follow-up independent samples t-tests were run to explore how boys (vs. girls) rated boy and girl targets in the intervention condition. Boys ($M = 3.71$, $SD = .52$) and girls ($M = 3.80$, $SD = .62$) did not rate boy targets differently in the intervention condition, $t(95) = -.79$, $p = .433$, Cohen's $d = -.16$. Boys ($M = 3.69$, $SD = .55$), in comparison to girls ($M = 4.04$, $SD = .63$), rated girl targets less positively in the intervention condition, $t(95) = -2.96$, $p = .004$, Cohen's $d = -.60$. Next, we ran paired-samples t-tests to explore how girls and boys rated girl (vs. boy) targets in the intervention condition. In the intervention condition, boys did not differ in how they rated the boy ($M = 3.71$, $SD = .52$) and girl ($M = 3.69$, $SD = .55$) targets, $t(47) = .41$, $p = .685$, Cohen's $d = .06$, and as noted in the prior paragraph, girls rated the boy targets less positively than the girl targets.

We also ran two univariate ANOVAs examining ratings of girl and boy targets to explore whether the Condition × Child Gender effect held among girl targets and boy targets, respectively. In these models, child gender and condition were included as fixed factors. Although the Condition × Child Gender effect was not significant for boy targets, $F(1, 182) = .73$, $p = .396$, $\eta_p^2 < .01$, it was significant for girl targets, $F(1, 182) = 6.73$, $p = .010$, $\eta_p^2 = .04$. Further follow-up independent samples t-tests explored how girls (vs. boys) rated girl targets in the control and intervention conditions, respectively. In the control condition, boys ($M = 3.95$, $SD = .62$) did not differ from girls ($M = 3.83$, $SD = .70$) in how they rated girl targets, $t(87) = .86$, $p = .390$, Cohen's $d = .18$, and as noted previously, boys (vs. girls) rated girl targets less positively in the intervention condition.

There was also a significant Target Gender Expression × Child Gender interaction effect, $F(1, 182) = 7.21$, $p = .008$, $\eta_p^2 = .04$ (see Figure A1 b). An independent samples t-test examined how boys (vs. girls) differed in their ratings of the gender-conforming and gender-nonconforming targets, respectively. An independent samples t-test showed that boys ($M = 3.83$, $SD = .68$) did not differ from girls ($M = 3.77$, $SD = .74$) in how they rated the gender-conforming target children, $t(184) = .54$, $p = .591$, Cohen's $d = .08$. Boys ($M = 3.79$, $SD = .61$) and girls ($M = 4.00$, $SD = .67$) also did not differ in their ratings of gender-nonconforming target children, $t(184) = -2.17$, $p = .032$, Cohen's $d = -.32$. Next, we conducted paired samples t-tests to examine how ratings of the gender-conforming vs. gender-nonconforming targets differed in boys and girls, respectively. Boys did not differ in their ratings of gender-conforming ($M = 3.83$, $SD = .68$) and gender-nonconforming targets, ($M = 3.79$, $SD = .61$), $t(91) = .49$, $p = .628$ Cohen's $d = .05$. In contrast, girls displayed lower ratings for gender-conforming ($M = 3.77$, $SD = .74$) than gender-nonconforming targets (M 3.99, $SD = .67$), $t(93) = -3.23$, $p = .002$, Cohen's $d = -.33$.

Additional Gender Stereotyping Findings

There was a significant Stereotype Category × Child Gender interaction, $F(1, 182) = 5.61$, $p = .019$, $\eta_p^2 = .03$ (see Figure A1 c). Independent samples t-tests were used to explore how the gender stereotyping of boys vs. girls differed on each of the masculine and feminine scales. Boys ($M = .28$, $SD = .34$), relative to girls

($M = .18$, $SD = .29$), displayed higher levels of gender stereotyping on the masculine scale, $t(178.37) = 2.33$, $p = .021$, Cohen's $d = .34$. In contrast, boys ($M = .22$, $SD = .28$) did not differ from girls ($M = .16$, $SD = .27$) on their level of gender stereotyping on the feminine scale, $t(184) = 1.39$, $p = .167$, Cohen's $d = .20$. Next, we split the file by child gender and ran paired-samples *t*-tests to examine ratings on the masculine vs. feminine scales within each gender group. Boys showed higher levels of gender stereotyping on the masculine ($M = .28$, $SD = .34$) than feminine ($M = .22$,

$SD = .28$) scale, $t(91) = -4.22$, $p < .001$, Cohen's $d = -.44$. In contrast, girls did not show different levels of stereotyping on the masculine ($M = .18$, $SD = .29$) vs. feminine ($M = .16$, $SD = .27$) scales, $t(93) = -1.11$, $p = .270$, Cohen's $d = -.12$.

Thus, although child participants' levels of gender stereotyping were relatively low in the present study (see Table 2), boys (vs. girls) displayed higher levels of gender stereotyping on the masculine scale, which measured how much children responded *only boys* to questions about activities stereotyped for boys. This finding is similar to those from Skočajić et al. (2020), who found that boys, but not girls, stereotyped masculine stimuli more so than feminine stimuli. Further, this finding aligns with previous literature that showed that boys (vs. girls) displayed more gender-conforming (i.e., stereotypically masculine) behavior (e.g., Spivey et al., 2018), and that self-perceived gender typicality was related to higher levels of gender stereotyping (Patterson, 2012). Our findings also suggest it is more important to boys than to girls to maintain masculine stereotypes, but not necessarily feminine ones (Halim et al., 2022). It is possible we found this result because boys more so than girls apply masculine stereotypes to themselves, and previous researchers have noted that stereotypically masculine traits are valued highly in North American society (Coyle et al., 2016). Thus, findings from the present study suggest that intervention work is still needed in Canada, especially for boys, to reduce their levels of masculine gender stereotyping.

Gender Stereotyping and Trait Empathy

The results of the mixed-model ANOVA for the gender stereotyping measure including trait empathy as a factor and modelling its interaction with other factors reported in the prior section are displayed in Table A1. As shown in Table A1, none of the main or interaction within-subjects effects were significant.

(a)

(b)

(c)
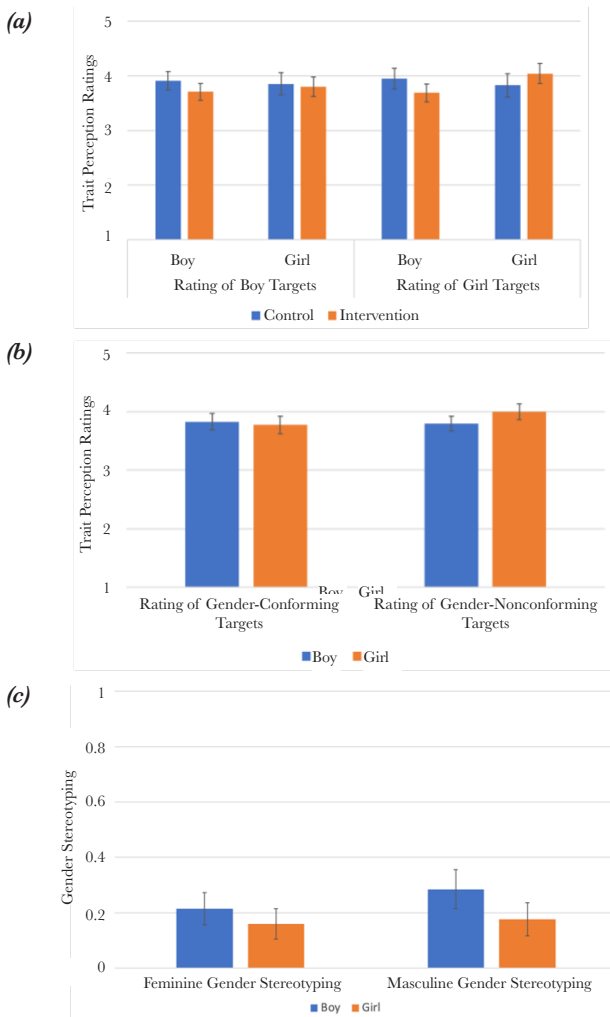
Figure A1. Trait Perception Ratings and Gender Stereotyping Ratings

Note. Error bars show 95% confidence intervals. Figure A1 (a) shows the Target Gender × Child Gender × Condition effect by plotting ratings of boy and girl targets among boy and girl participants per condition. Figure A1 (b) shows the Target Gender Expression × Child Gender effect by plotting ratings of gender-conforming and gender-nonconforming targets among boy and girl participants. In (a) and (b), higher scores represent more positive appraisals of the target children. Figure A1 (c) shows the Stereotype Category × Child Gender effect by plotting levels of feminine and masculine gender stereotyping among boy and girl participants. In Figure A1 (c), higher scores represent higher levels of gender stereotyping.

Table A1. Mixed-Model ANOVA for Trait Perception Ratings for Gender Stereotyping with Trait Empathy

	F-value	*p*	Partial Eta-Squared
Stereotype category	2.37	0.125	0.01
Stereotype Category × Condition	1.02	0.315	0.01
Stereotype Category × Child Gender	0.000	0.997	<.01
Stereotype Category × Empathy	1.11	0.293	0.01
Stereotype Category × Condition × Child Gender	0.32	0.573	<.01
Stereotype Category × Condition × Empathy	1.47	0.227	0.01
Stereotype Category × Child Gender × Empathy	0.08	0.778	<.01
Stereotype Category × Condition × Child Gender × Empathy	0.21	0.646	<.01
Condition	0.30	0.585	<.01
Child gender	0.24	0.623	<.01
Empathy	<.01	0.950	<.01
Condition × Child Gender	0.05	0.823	<.01
Condition × Empathy	0.53	0.467	<.01
Child Gender × Empathy	0.06	0.813	<.01
Condition × Child Gender × Empathy	<.01	0.999	<.01

Appendix C

Vignettes
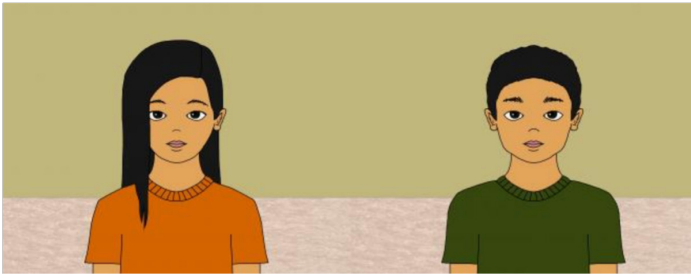
This is a girl named Olivia and a boy named Ethan. Olivia and Ethan are in the same grade as you.

(1) Olivia's favourite toy is her Spiderman and Ethan's favourite toy is his tea set. Olivia's favourite activity is hockey and Ethan's favourite activity is making jewellery. At school, Olivia and Ethan get teased and excluded by their peers. Olivia's peers tease her because they think Olivia likes toys and activities that are for boys, and Ethan's peers tease him because they think he likes toys and activities that are for girls.

(2) Olivia wants to be a pilot and Ethan wants to be an elementary school teacher in the future. Olivia's favourite subject at school is Math and Ethan's favourite subject is Reading. At school, Olivia and Ethan are told by their teacher that they should pick different careers and favourite classes. Their teacher says that boys, and not girls, should be pilots and like Math. And their teacher says that girls, and not boys, should be elementary school teachers and like Reading.

(3) Olivia likes to wear a firefighter costume and Ethan likes to wear a cheerleader costume during dress-up. Olivia's favourite chore is to take out the garbage and Ethan's favourite chore is to set the table for dinner. At home, Olivia's parents and Ethan's parents do not allow them to wear the clothes or do the chores they like. Olivia's parents say that girls should not wear firefighter clothes and take out the garbage. And Ethan's parents say that boys should not wear cheerleader costumes and set the table for dinner.