PNAS
www.pnas.org

Supplementary Information Appendix for

Similarity in Transgender and Cisgender Children’s Gender Development

Selin Gülgöz, Jessica J. Glazier, Elizabeth A. Enright, Daniel J. Alonso, Lily J. Durwood, Anne A. Fast, Riley Lowe, Chonghui Ji, Jeffrey Heer, Carol Lynn Martin, Kristina R. Olson

Selin Gülgöz
Email: sgulgoz@fordham.edu

This PDF file includes:

Supplementary Information Text

1. Registration

The following is a revised copy of the official registration. We wrote the original registration (registered online on 01/22/2018 at https://osf.io/q2kuw/?view_only=c9f9df7d1e5a4f95ab893f28a81ce9e0) after data had been collected but before analysis for this paper. Some findings with smaller subsets of participants—particularly main effects on the primary variables—had previously been published (*1-4*) and were therefore known to the authorship team. For this reason, this registration should not be confused with a *pre*registration. Our primary goals in registering this study were to (a) pre-define how variables would be coded to reduce researcher degrees of freedom and (b) decide which analyses we planned to run to reduce concerns that a discovery in data analysis would lead to another analysis that could be misremembered as pre-planned. Because the registration was written and published online before we examined the data, we noticed later on that some information was incorrect. Below, we include a revised version of the registration both showing the original information (shown in strikethrough) and reflecting the changes that were made (shown in bold).

In the main text of this study, we reported a subset of the registered analyses, due to space constraints. Specifically, we reported analyses relating to child measures that include participant group comparisons, comparisons to chance within each participant group, coherence between measures within each participant group, and changes related to time since transition for transgender participants. All other registered analyses, listed below, are included in this supplement.

The registration, as posted online on 01/22/2018, is as follows:

Data collection and participants:

- This project is part of a larger, longitudinal study, the TransYouth Project.
- Participants are transgender children, their gender typical siblings, and gender typical controls who are matched by age and gender to each transgender participant. Transgender participants are defined as those participants who use the pronouns opposite their sex at birth in all contexts (i.e., school, in public, with all parents, etc.).
- All participants are between ages 3 and 12 years.
- Data collection from transgender children and their siblings took place between July 2013 and December 2017. Matched controls will continue until they have all been run (anticipated end of February 2018). Exact participant numbers will not be known until data collection is complete, and will vary by measure as some participants will not have completed all measures listed below. Every transgender child aged 3-12 run during this time will be included, with the following exceptions: children who had low cognitive or verbal capacity as demonstrated by inability to answer questions (the study is meant to be about children without major developmental delays), children who did not understand English (though this has not yet occurred at the time of registration), children who have previously participated in the study (e.g., Visit 2 of ongoing longitudinal study). In addition, control participants (i.e., gender-typical participants matched by age and gender to each transgender participant) will be excluded for the same reasons. For each transgender participant, we collected data from one gender-typical sibling (the closest in age between 3-12 years of age, when possible), and in cases where a sibling had a major developmental impairment (e.g., low cognitive or verbal capacity) or was unavailable during the visit we recruited the next closest in age sibling in that age range.
- Throughout this registration, the term 'gender' is used to refer to the gender associated with the child's gender pronouns in everyday life. This means that for siblings and controls, this is the gender that aligns with their sex at birth, and for transgender children this is the gender opposite their sex at birth.
- Partial data from three subsets of participants included in this ~~monograph~~ **paper** were reported in published papers **before this registration was written (*1-4*~~)~~ 2018; see Table S1 in this supplement for numbers of overlapping participants)**. In these papers there were ~~36~~ **31**, 32,

and ~~70~~ **31** transgender children respectively (some of the N=70 were also part of the N=32**; the original "70" in the registration was an error as this included children who did not meet the criteria of "transgender"**), and often on only a subset of the present measures. The present paper will report on all data from Visit 1 of **more than** 300 transgender children in comparison. Across the paper and the supplement, we will report the results for each measure both including and excluding the previously reported data to assure maximum transparency.

Measures and scoring:

- Child measures (DVs):
  - Toy preferences: Toy preference scores will be computed by recoding items on each trial so that items most associated with a participant's gender is coded as a 5, those least associated with a participant's gender (i.e., most associated with the opposite gender) will be coded as 1, and scores 2 through 4 will represent the spectrum between. The items completed will be averaged, and scores will be recoded on a 0-100 scale, such that 100 means a participant consistently selected items most associated with their own gender, and 0 means a participant consistently selected items most associated with the opposite gender. Participants will receive a score on for this task only if they complete at least 50% of items (2 of 4 trials).
  - Clothing preferences: We will use the same scoring as with the toy preferences task.
  - Peer preferences: Peer preferences will be computed as the percentage of time children selected own-gender targets (e.g., percentage of times a girl selected girls across trials completed), where a score of 100 would indicate strong same-gender preferences, and a score of 0 would indicate strong opposite-gender preferences in peers. Children will be included in this measure if they complete at least 50% of the items (3 of 6 trials).
  - Gender identity Implicit Association Test (IAT): Scores will be calculated using the d-score algorithm. Responses will be coded such that higher positive scores indicate higher implicit identification with one's gender and lower negative scores will indicate higher implicit identification with the opposite gender, with scores around 0 indicating equal association with both genders. Children below age 6 and those who cannot read were not asked to complete this measure. In addition, children making errors on more than 30% of trials, or who complete more than 10% of their responses in less than 300ms will be excluded.
  - Gender identity now and predicted gender identity as an adult: Scores will be calculated such that participants receive 1 point for responding with their identified/current gender, -1 point for responding with the opposite gender, and 0 points for an "other" response (including response choices: 'neither', 'both', 'it changes over time', 'I don't know'). For certain analyses (described in research questions #6-9) these two items will be added together for an overall explicit identity composite (resulting in a score from -2 to +2). Participants must have responded to the item to be included in individual analyses and must have responded to both items to be included in the composite.
  - Similarity: Scores will be calculated separately for similarity to own gender and similarity to the opposite gender, averaging all items about each gender into a single composite. Scores will range from 1 (totally different) to 5 (totally the same) for each. For certain analyses (see research questions #6, 7, 9), we will use a difference score of (similarity to own gender) – (similarity to opposite gender), resulting in a scale from -4 to +4. Participants will be excluded from a composite if they didn't respond to at least 3 items within each 5-item composite and difference scores will not be computed if they didn't respond to at least 3 items on each composite.
- Additional dependent variable:
  - Outfit at appointment: Outfits will be scored from 1 (highly stereotypically masculine) to 5 (highly stereotypically feminine) by two independent researchers. The two raters' scores will be averaged for each participant, and converted to a 1 (highly stereotypical of

opposite gender) to 5 (highly stereotypical of own gender) scale. In cases where only one coder was available, that score will be used. If experimenters failed to score children's clothing, children will be excluded from this measure.

- Parent measures:
  - Demographics (child's race, household income, parental political orientation, parent education level, number of siblings, geographic location): When two parents completed demographics, mothers' responses will be utilized if available. If no mother was present or if two mothers were present, the parent who was the primary study contact's responses will be utilized. Race will be recoded as White and Not-White (the latter including multi-racial children for whom White is one race), and used as a categorical variable. Household income, parental political orientation, parental education, and number of siblings will be used as continuous variables in analyses. Geographic location will be treated as a categorical variable with categories defined at end of document.
  - Parent report on child's gender identity now and in adulthood: These responses will be scored identically to the child measures of explicit gender identity (1 point for responding with child's identified gender, -1 points for responding with the opposite gender, 0 points for other responses). For participants with two parents participating in the study, we will average the parents' responses. For analyses described in research question #8, we will use a composite score of the two items, adding them together for an overall explicit identity composite (resulting in a score from -2 to +2, similar to the composite in the child measure).
  - Parent report on child's behaviors and preferences: Parents are asked to answer 8 questions asking their child's preferences (e.g., in toys, clothing, video game avatars) and are given the options to respond that their child's preferences are more typical of boys, more typical of girls, or gender neutral. Responses will be scored as 1 point for child's gender, -1 points for the opposite gender, and 0 points for gender neutral responses, or if parents circle both genders. If parents circle one of the gender and the neutral option, we will average the scores. For participants with two parents participating in the study, we will average the parents' responses.
  - Gender Identity Questionnaire (GIQC; Johnson et al., 2004): Scoring will be done in accordance with the method described in Johnson et al. (2004), including reverse scoring some items and dropping the two items they dropped (as well as the two extra items we added) for computing the composite (however, we will include means per item, for all items, by participant group in a table for interested parties). For comparison to past work, and in accordance with Johnson et al, responses to each item are coded from 1 to 5, where 5 indicates responses most aligned with one's sex at birth (meaning scores of 1 are most aligned with transgender children's current gender and are aligned with the opposite sex for control and sibling groups). These scores will be reported for easy comparison to past work. However, for primary analyses throughout the paper, scores will be reversed for transgender participants so that this measure aligns with all other measures in this paper, which are coded according to gender identification and not sex at birth. On some questions, parents are also given the option to respond that the question does not apply to their child, in which case those items will not be included in the calculation of the average (as per Johnson et al.), and the denominator will be modified accordingly. If parents did not complete this measure they will not be included in analyses including this measure. For participants with two parents participating in the study, we will average the parents' responses and report the overall correlation between the two parents' responses.

Research questions, planned analyses, and hypotheses:

To address each research question (below), we will first include the descriptive statistics (means, standard deviations by participant group, age, and gender) of each measure, then compare participant groups via

ANOVAs with participant group (transgender, controls, sibs) and gender as between-subjects factors, following up significant effects with post-hoc Tukey comparisons, separately for each dependent measure.

In addition, we will compare responses to the neutral/chance responding via one-sample *t*-tests. If there are no significant participant group differences (or an interaction) found in the ANOVA, we will conduct these *t*-tests for the whole sample, collapsing across participant group for a given DV. Similarly, if there are no significant differences based on gender, we will collapse across gender for the *t*-tests on that DV. If there are differences by participant group, gender, or an interaction for a given DV, follow-up one-sample *t*-tests will be conducted within those subgroups.

In addition, we will conduct Pearson correlations to assess the relation between age and each of the DVs.

For research questions that require exceptions to this analysis plan, we describe each specific plan below the relevant question.

In general, because we will have considerable statistical power, and therefore traditional significance (p<.05) can be more easily achieved, we will focus our interpretation on differences between groups with meaningful effect sizes (defined as d>.20, r>.10). Negligible but significant effects will be mentioned but not emphasized.

12. Are there differences between transgender children, their siblings, and matched controls in terms of their toy, peer, or clothing preferences? Will children show own-gender preferences on these measures?

Hypothesis 1a: We do not predict any participant group or gender differences with meaningful effect sizes.

Hypothesis 1b: We predict that children will show strong preferences for toys, peers, and clothing associated with their gender.

13. Are there differences between transgender children, their siblings, and matched controls in terms of their IAT scores? Do children implicitly associate themselves with their gender identity?

Hypothesis 2a: We do not predict any group or gender differences with meaningful effect size.

Hypothesis 2b: We predict that children will significantly associate themselves with their gender identity on the IAT.

14. Are there differences between transgender children, their siblings, and matched controls in terms of their explicit gender identity now and their predicted gender identity in the future? Do children explicitly identify as their own gender?
    a. Separate chi-square tests will be conducted for the now and the future items separately. If significant, follow-up chi-square tests will be computed within pairs of groups to determine which groups differ from one another.
    b. We will compute chi-square goodness of fit tests to assess whether children were more likely than chance (calculated in comparison to 33/33/33% probability) to select their own gender.

Hypothesis 3a: We do not predict any group differences or gender differences with meaningful effect size (φ = .10).

Hypothesis 3b: We predict that the children in every group will be more likely than chance to identify as their gender and that, in fact, the majority of each group will do so both when asked about now and in the future.

15. Are there differences between transgender children, their siblings, and matched controls in terms of their perceived similarity to their own and opposite gender? Do children view themselves as more similar to their own gender, than the opposite gender?

Hypothesis 4a: We do not expect any group or gender differences with meaningful effect size on perceived similarity to own vs. opposite gender.

Hypothesis 4b: We predict that children will significantly associate themselves with their gender (one-sample comparison to 0 for the difference score).

16. Are there differences between transgender children, their siblings, and matched controls in terms of how gender-stereotypical their outfit was at time of appointment? Do children wear clothing that is more stereotypical of their own gender?

Hypothesis 5a: We do not expect any group or gender differences with meaningful effect size in terms of how gendered children's outfits are at time of appointment.

Hypothesis 5b: We predict that children dress in line with gender stereotypes.

17. Do children in different groups show coherence in their responses across the different types of measures?
    a. Correlation analyses will be conducted separately for each participant group, between all DVs listed above.

Hypothesis 6: We expect child measures to show coherence (i.e., correlations of $r > .20$ between all measures) for all three groups of participants.

Secondary questions and exploratory analyses:

The research questions in this section are exploratory in nature, for which we do not have any clear predictions. For this reason, although we will report all findings as they will be important for informing future work, we will focus our interpretation on findings that are significant at the level of $p < .005$.

18. Do participants' demographics relate to their responses to the different measures?
    a. Detailed tables will show demographic distribution of participants in terms of race, geographical location, political ideology, parent education level as well as how each of these responses is related to means on each DV (i.e., the mean toy preference for White and non-White children, the mean peer preference for children in the Pacific Northwest, Northeast, etc.).
    b. An independent sample t-test will be conducted on each DV by race (white vs non-white) within each participant group. When differences are significant, a separate one-sample t-test will be conducted within the White and nonwhite groups to assess whether the overall effect was significant (e.g., whether White children show a significant preference for gender-stereotypical toys). If there is no race effect, no further analyses will be conducted.
    c. A one-way ANOVA with geographic location (6 pre-determined groups described below) as a between-subjects factor will be conducted for each DV, once for transgender participants and once for siblings. If significant, Tukey post-hoc analyses will be conducted to determine which locations differ from which other locations. Further, if regions differ, separate one-sample t-tests will be conducted for the relevant DV to assess whether the overall effect is significant within each region (i.e., children in the Pacific Northwest show significant gender-stereotypical toy preferences). Because control participants are all from one geographic region, these analyses will not be computed for them.
    d. Correlation analyses will be conducted between three demographic factors (parent education level, parent political ideology, household income) and DVs within each participant group.
19. How do parental reports of transgender children's gendered preferences and behaviors relate to children's gender identity? Do parents' identification of their children's gender identity correlate with children's self-reported identities?
    a. For all analyses involving the GIQC, we will remove the items excluded by Johnson et al. and, because the transgender children are socially-transitioned, all children will be scored according to gender (not sex at birth). This means the scoring procedure Johnson et al. used for controls will be used for all of our participants). First, we will assess whether groups and genders differ on this measure by computing an ANOVA with participant group and gender as IVs, and mean score on the GIQC as DV, in line with our general analysis plan. Again, we will use Tukey tests for follow-up comparisons if significant effects or an interaction occurs.

b. Scores on the GIQC will also be correlated with the dependent variables, separately for each participant group.

c. Separate correlation analyses for each participant group will be conducted between parents' reports of children's gender identities (using the composite score adding gender identity now and future predicted gender identity) and children's own implicit and explicit reports of their gender identities.

d. Separate correlation analyses for each participant group will be conducted between parents' reports of children's toy, clothing and peer preferences, and children's scores on each of the parallel tasks.

20. Are there developmental changes in participants' scores on each of the measures?

    a. Age correlations will be conducted with each child DV listed above, for each participant group, in addition to showing the general patterns by age in a table.

    b. We will also conduct partial correlations between time since transition and each child DV listed above, controlling for age of participant. These analyses will be conducted only for transgender participants.

Groups of geographical regions:

1. Northeast: CT, MA, MD, ME, NH, RI, VT, NJ, NY, PA, DC, DE, ON (Canada)
2. Midwest/Upper Plains: IL, IN, MI, OH, WI, IA, KS, MN, MO, NE, ND, SD
3. Southeast: FL, GA, NC, SC, VA, WV, AL, KY, MS, TN, AR, LA, OK, TX
4. Mountain West: AZ, CO, ID, NM, MT, UT, NV, WY
5. Pacific NW: OR, WA, BC (Canada), AK
6. Pacific South: CA, HI

2. Detailed Method

Participants

All participants completed the current study during their first visit of what is an ongoing longitudinal study investigating gender cognition among three groups of participants: socially-transitioned transgender children, cisgender siblings of transgender participants, and cisgender controls who were matched in age and gender to each transgender participant. For each group of participants, data were also collected from at least one (and when present, two) parent(s). Participants completed the current study between 07/2013 and 02/2018 (transgender participants and siblings were recruited between 07/2013-12/2017; control participants were recruited between 07/2013-02/2018 due to matching procedures), at which time each child was between 3 and 12 years old. Before beginning the study, parents provided verbal and written consent. Additionally, children ages 3 to 8 provided verbal assent, and children ages 9 to 12 provided verbal and written assent. During consent and assent processes, participants were told about the study, had the opportunity to ask questions, and were told they could skip any questions they wanted to or could quit the study at any time without loss of remuneration ($10 per child and per parent plus a small toy for each child). Below, we describe recruitment methods and characteristics unique to each participant group.

Transgender children. 317 transgender participants (*M* = 7.62 years old, *SD* = 2.37) were included in the current sample. Transgender participants had already socially transitioned by the time they participated in this study. This means that, across every context (at home, at school, meeting new people), they were using a binary pronoun (i.e., he or she) that was not associated with their sex assigned at birth. Of the transgender children, 208 were transgender girls (assigned males) and 109 were transgender boys (assigned females). A preponderance of assigned males is often seen in studies of gender diverse children in early childhood *(e.g., 5-6)*.

Transgender participants were recruited across the U.S. and Canada through support groups, conferences and summer camps for gender nonconforming children and their families, word-of-mouth, online through our project website, or in response to media coverage of our ongoing research (see Fig. S1 for a map of where participants are located). Some participants were also recruited through our process of recruiting controls, when families who were invited to participate as control participants stated that they have a transgender child. Interested participants were asked to sign up for the study online on our project website. Once a group of families within a geographical region had signed up for the study, a team of 2-3 researchers traveled to where the families were located. Our research team conducted trips approximately once a month during the testing period. Participants were met in their homes, in private rooms arranged in public libraries or universities, or in private spaces at aforementioned conferences for families with gender nonconforming children. Some transgender participants (and their siblings and parents) participated in a developmental psychology lab space. Each participant's session lasted approximately 30 minutes.

Full demographics for this sample can be seen in Table S1. Whereas the sample included people from many demographic groups, the families were, on average, wealthier, more liberal, more likely to be White, and more educated than the national average ("US Census Bureau, 2017"). These characteristics are often true of families who participate in research studies *(7)*; thus, it is difficult to determine the degree to which these characteristics are associated with having a socially-transitioned transgender child in the 2010's, or if they are simply the kinds of families who have the time, trust, and interest to participate in a university-based research study. More families signed up than could be accommodated. Therefore, the research team ran as many families as possible given constraints such as money (some families were in locations that were too expensive to visit), time (if one family was many hours' drive from any other family we sometimes opted to run several families near one another instead due to limited time for travel), and the availability of families (sometimes they were out of town or the child got sick when we were in their area and therefore they were not able to participate). When resources were limited and we could not reach all families in a given area, and when we had such information in advance, we preferentially included non-White children and families whose annual household income was less than

$75,000 to increase the representation of these groups in our sample (who tended to be underrepresented compared to the U.S. population).

As has been reported in other samples of gender nonconforming and transgender children *(5, 8)*, adopted children were overrepresented in this sample (7.3% adopted compared to an estimated 2.5% of children adopted in the U.S.; "U.S. Census Special Reports"), though adoption rates are drastically lower than in some other papers (see *(5)* for a sample of >50% adopted children). Wealthy and more educated families are more likely to adopt children *(9)*, so again, whether this is uniquely true for samples with transgender children and/or the kinds of families who sign up for research studies is currently unknown.

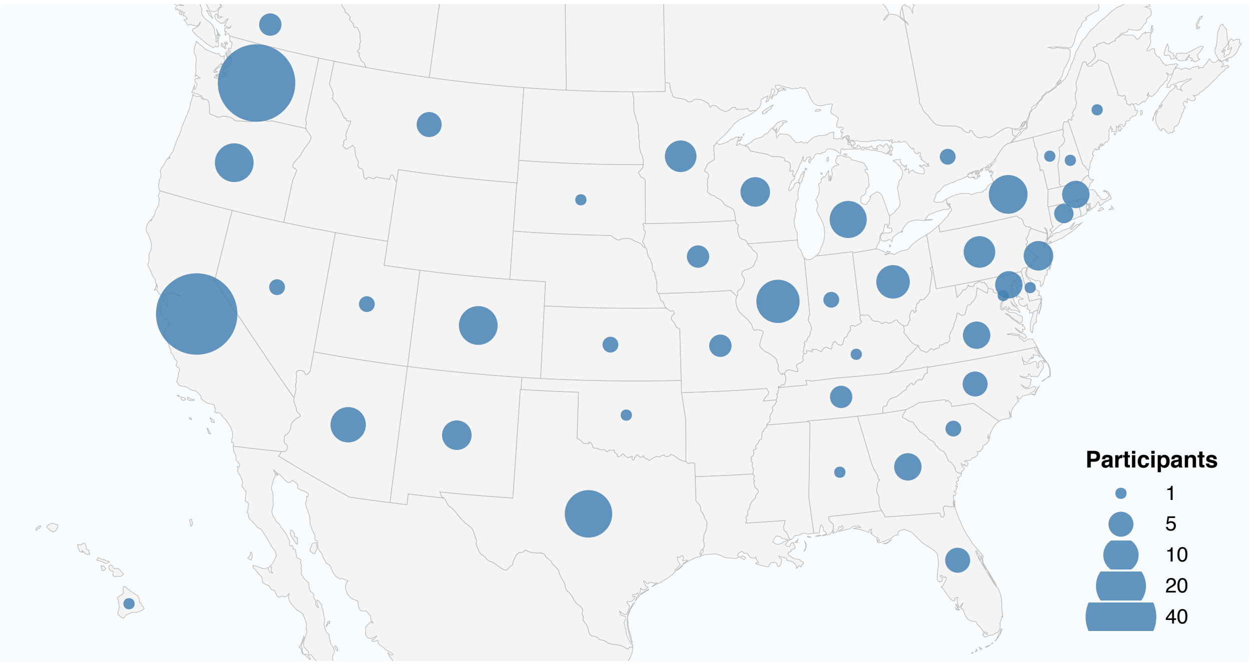

Fig. S1. Map of states where transgender participants (and their siblings) reside. Larger circles indicate greater numbers of participants.

Table S1. Participant Family Demographics[1]

	Controls	Transgender	Siblings	Statistic
Child's Race[2]				$\chi^2(2)=.14$, $p = .933$
White: Non-Hispanic/Latino	70%	68%	68%	
Hispanic/Latino	1%	2%	4%	
Black	1%	2%	2%	
Asian	2%	3%	2%	
Native American	1%	<1%	1%	
Asian/White	12%	5%	5%	
Latino/White	5%	7%	1%	
Black/White	2%	4%	2%	
Other Multiracial	7%	9%	7%	
Not Reported	1%	1%	2%	
Household Annual Income[3]				$F(2,812)=11.58$, $p <.001$, $\eta_p^2 = .03$
Less than $25,000	2%	4%	5%	
$25,001 to $50,000	5%	10%	10%	
$50,001 to $75,000	12%	20%	18%	
$75,001 to $125,000	31%	31%	31%	
More than $125,000	48%	35%	35%	
Not Reported	1%	1%	1%	
Parent Education Level[4]				$F(2,303)=.43$, $p = .650$, $\eta_p^2 < .01$
Some schooling	0%	0%	0%	
High school diploma	0%	1%	2%	
Some college/Associate's degree	14%	18%	21%	
College/Bachelor's degree	48%	33%	35%	
Advanced degree (MA, MD, PHD, etc.)	38%	49%	43%	
Parent Political Ideology[5]				$F(2,813)=44.11$, $p < .001$, $\eta_p^2 = .10$
Liberal	63%	86%	84%	
Moderate	33%	11%	13%	
Conservative	3%	1%	1%	
Not Reported	1%	<1%	1%	

Geographic Location[6]			
Northeast		13%	15%
Midwest/Upper Plains		21%	24%
Southeast		15%	15%
Mountain West		13%	15%
Pacific Northwest	100%	20%	19%
Pacific South		16%	12%

Note. Although occasionally two parents completed the demographics form, responses from one parent (the primary contact) are reported above. All statistical comparisons are between the control and transgender groups.

[1] Percentage are reported in terms of total participants.

[2] Percentage of child in each racial category. Chi-square analysis was conducted with two categories (White and Non-White).

[3] Percentage of families in each income bracket. One-way ANOVA analysis was conducted on a 5-point scale variable that maps onto the five options provided. Note that two control families circled both $50,001-$75,000 and $75,001-$125,000; these families were not included in this table. In our analyses, these families' income scores were coded as "3.5" on the scale.

[4] The parent education level item was added later than other demographic items. Therefore, data are missing on this item for 62.3% of cisgender control participants, 62.5% of transgender participants, and 64.0% of the cisgender sibling participants. Missing data was not included when calculating the percentages listed in this table for this item.

[5] Mean and standard deviation of parents' political ideology on a scale ranging from (1) very liberal to (7) very conservative.

[6] 1.3% of the parents of cisgender controls did not respond to this item. However, we have counted them as being located in the Pacific Northwest (here and in subsequent tables and analyses) since all control participants were recruited and run in that area of the country. The geographical regions were defined in the following manner: Northeast = CT, MA, ME, NH, VT, NJ, NY, PA, DC, DE, ON (Canada); Midwest/Upper Plains = IL, IN, MI, OH, WI, IA, KS, MN, NE, SD; Southeast = FL, GA, MD, NC, SC, VA, WV, AL, KY, MS, TN, LA, OK, TX; Mountain West = AZ, CO, NM, MT, UT, NV, WY; Pacific Northwest = OR, WA, AK, BC (Canada); Pacific South = CA, HI. No participants were from RI, ND, MO, AR, or ID.

Cisgender siblings of transgender children. Siblings of transgender participants were recruited through the same mechanisms as the transgender participants. If the transgender participant had a sibling, the sibling closest in age to each transgender participant, who was between the ages of 3 to 12 years old, was also invited to participate. When the closest-in-age sibling was unavailable (i.e., was out of town) or an exclusion criterion applied (i.e., major developmental delay), the next closest in age sibling participated, if one was available. Of the 317 transgender participants, 189 had a sibling participate (107 brothers, 82 sisters; *M* = 7.61 years old; *SD* = 2.45 years) and 128 did not have a participating sibling. Additional demographic information is included in Table S1.

Gender- and age-matched controls. We also recruited a control group of cisgender participants to match each transgender child by gender and age. For example, once a 7-year-old transgender girl (assigned male) participated, we would recruit a 7-year-old cisgender girl as a control. On the day they were tested, all control participants were within four months of age of the transgender matched participant on the day that child was tested. Control participants were recruited from a university database of families who indicated interest in participating in research at the time their child was born, and therefore were all from one geographic area—a major metropolitan area in the Pacific Northwest. The demographics of this geographic area, including the political orientation (politically liberal), racial make-up (largely White), and average household income (high income) were relatively well-matched to the transgender and sibling comparison groups, though some differences occurred (see Table S1). Control participants were tested in a research lab. Before agreeing to participate in the study, parents were informed that this was a longitudinal study investigating gender development in transgender children and were told that the study involves explaining to participants that some children identify as the gender "opposite" their assigned sex. Given that study participation is optional, only parents who were comfortable with the topic of the study participated. If any parents reported their child was transgender, they were moved to the transgender group (or if they were gender nonconforming but had not transitioned, they were excluded from this study). In a few cases the parent stated that the child was the sibling of a transgender child in which case the transgender child was asked to come in as well and the originally-recruited child was then considered a sibling rather than an unrelated control. During the testing period, if participants were recruited for another study by the same laboratory and a parent mentioned they had a transgender child, that family was asked if they wanted to participate in this study. For these reasons, a disproportionate number of families of transgender children came from the researchers' regional area. A total of 316 control children participated (*M* = 7.66 years, *SD* = 2.40), including 207 girls and 109 boys to match the transgender participants.

Measures and procedure

Children were presented with questions about their toy preferences, peer preferences, clothing preferences, their current gender identity, the gender identity they thought they would be as an adult, and how similar they felt to other boys and girls. These questions were asked verbally, and children could respond by answering the question out loud or pointing to different options on the computer or response sheets. Children were also asked to complete a Gender Identity Implicit Association Test (IAT), and researchers rated how gendered the child's outfit was at the appointment. Each of these measures is described in further detail below. Small subsets of children in this study also completed other tasks (e.g., 106 children completed a gender stereotyping measure), but the current measures are included in this report because they were given to all or nearly all participants and are relevant to gender development and consistent of the standard battery given to these participants in the first year as well as later years (while other measures vary). Details about requirements for inclusion are described below for each measure. For each task, Table S2 shows the number of participants who received the task and the number of participants that completed it, the number of participants who did not receive the task and reasons why, and the number of participants whose data were described in previous published work.

Table S2. Details Regarding Number of Participants Reported for Child Tasks

Task	Received Task	Completed Task	Did Not Receive Task	Reasons for Not Receiving Task			Data Described in Previous Work		
				Task did not exist at time of participation	Incorrect age for task	Experimenter Error	Olson et al. (2015)	Fast & Olson (2018)[1]	Rae & Olson (2018)[1]
Toy Preferences	720	712	102	53	46	3	0	76	0
Clothing Preferences	720	712	102	53	46	3	0	76	0
Peer Preferences	773	749	49	0	46	3	77	75	0
Similarity	720	689	102	53	46	3	0	61	0
Gender Identity IAT	582	552[2]	225[3]	0	180	0	69	0	68
Current Gender Identity	822	810	0	0	0	0	78	74	0
Future Gender Identity	822	816	0	0	0	0	77	72	0
Outfit at Appointment	768	766	56	53	0	0	0	74	35

[1] Some of the participants in the Fast & Olson (2018) and Rae & Olson (2018) papers were also in the Olson et al (2015) paper so these numbers are not mutually exclusive.

[2] 14 participants who completed the IAT were excluded from IAT analyses for making errors on more than 30% of trials and/or completing more than 10% of their responses in less than 300 ms.

[3] 3 participants did not receive the IAT for unknown reasons and the experimenter failed to write down the reason. 42 participants did not receive the task because they could not read.

Parents filled out a questionnaire packet on their own during the visit. Specifically, they were asked demographic information (see Table S1) and questions about their child's gendered behaviors. In 72% of cases, transgender children and their siblings had two parents fill out the questionnaire. The control children only had one parent fill out the questionnaire because only one parent was present during the assessment. Parents filled out additional measures beyond those reported here (e.g., mental health questions, questions about support for their child's gender identity), but only the questions relevant to the present research questions are reported in this paper.

We registered our measures, exclusion criteria, scoring, and analyses prior to conducting any analyses on the complete data https://osf.io/q2kuw/?view_only=c9f9df7d1e5a4f95ab893f28a81ce9e0) and all details below regarding exclusion, scoring and analyses align with that registration unless otherwise noted.

Child measures

Toy preferences. To determine children's toy preferences, we asked children to indicate the toy they would like to play with the most from four arrays of five toys ranging from stereotypically "boy" toys to stereotypically "girl" toys. The toys participants saw were dispersed throughout the slide so that the pictures were not in order by their stereotypic gender rating. In cases where participants selected two or more toys, the experimenter prompted them to choose their one favorite.

To make sure toys were age-appropriate and attractive to participants, two sets of stimuli (pictures of toys) were created, one for younger children (ages 3-7) and one for older children (ages 8-11). The toys included in the trials were all pilot tested with a separate group of cisgender children to determine which toys were perceived by children (from each age group) as more or less feminine or masculine. Sets of five were created such that each set had a highly masculine, a moderately masculine, a neutral, a moderately feminine, and a strongly masculine toy.

This measure was scored such that children received higher scores for choosing toys most consistent with stereotypes of their own gender. So, if a girl chose the most stereotypic "girl" toy (e.g., a baby doll), she would receive a score of five, but if a boy chose that same toy, he would receive a score of one, because it was the least stereotypic "boy" toy. Children's scores were averaged across the four arrays and converted to a 0-100 scale such that 100 meant a participant consistently selected the most stereotypic toys associated with their gender on every trial, and 0 meant a participant consistently selected the most stereotypic toys associated with the opposite gender (Cronbach's α = .63). Participants had to complete at least 50% of trials on this task (i.e., two out of four) for their data to be included in the final sample for this measure.

Clothing preferences. To determine the degree of children's gendered clothing preferences, children were given a clothing measure modeled exactly on the toy measure (Cronbach's α = .69). All aspects of piloting, presentation order, and scoring were identical to the toy preference measure except that the same items were used for all ages.

Peer preferences. To determine whether children preferred same-gender or opposite-gender peers, we presented children with six boy-girl pairs and asked children to choose who they would like to be friends with the most (two filler trials – one with two boys and one with two girls were included but did not contribute to scoring). For half of the trials the boy was on the left, and for half of the trials the boy was on the right. Six trials (five real and one filler) included two White children, one real trial included Black children, and one filler trial included Asian children. In this way, the trials reflected some racial variation, but most target pictures were White to reflect the fact that we anticipated most participants would be White, but there would be some variability on race.

The measure was scored by counting the number of times children chose a same-gender playmate on the six mixed-gender trials, dividing by the number of mixed-gender trials they completed, and multiplying the result by 100 to again produce a value from 0 to 100, where 100 indicated strong same-gender preferences and 0 meant strong opposite-gender preferences. Participants had to complete at least 50% of trials (at least three of six mix-gender trials) for their data to be included in the final sample of the current measure.

Current gender identity. Children were asked to indicate their current gender identity by selecting from a list of options. The first 416 participants were given a version of the task where they were first told that some people feel like they are a boy/girl on the inside and this could be the same or different from what they are on the outside. Then they were asked what they felt like on the inside and could choose that they felt like one of the following options: "a boy," "a girl," "neither," "both," "it changes over time," or "I don't know". 412 participants received a simplified version of the task, where they were not told about the different outside/inside feelings and were just asked what they feel like they are. They were given the options of "boy," "girl," or "something else". If they chose something else, they were then given the options of, "neither," "both," "it changes over time," or "I don't know". Importantly, control participants received the same version as their matched transgender participant.

Children received a score of 1 if they responded with their current gender (the gender aligning with their current pronouns), a score of -1 if they responded with the opposite gender, or 0 for any "something else" response (including, "neither," "both," "it changes over time," or "I don't know"). This meant that, if control participants and siblings indicated their assigned sex as their gender identity they received a 1, and if transgender participants indicated their current gender identity (opposite their assigned sex) they received a 1; an assigned sex answer led transgender participants to receive a score of -1.

Future gender identity. Children were also asked to predict their gender identity in the future. Response options were identical to those in the Current Gender Identity item and were scored identically.

For additional analyses correlating gender identity with other measures, scores from the current and future gender identity questions were added together for an overall explicit identity composite (resulting scores ranging from 2 to -2). We did this because the current and future gender identity questions were correlated, $r(808) = .484$, $p <. 001$, and doing so gave us a more reliable estimate of children's gender identity. Participants needed to have responded to both items to be included in this composite.

Similarity to boys and girls. To measure how similar children felt they were to girls and to boys, we used a measure from Martin and colleagues *(10)*. Children were shown a scale (see *10*), and were told they were going to answer a few questions about how similar they are to other kids. Children were given two scales, one that they would use to respond to questions about how similar they are to girls, and the other to use to respond to questions about how similar they are to boys. Children were told that if they responded with a '1' on either scale, it meant they thought they were totally different from girls/boys, a '2' indicated kind of different, a '3' indicated in the middle, a '4' indicated kind of the same, and '5' indicated totally the same as girls/boys.

Children were asked ten questions about how similar they felt to other kids. Five questions were about how similar they felt to girls and five questions were about how similar they felt to boys. The questions included global questions of how similar participants felt to girls/boys, how much they act like girls/boys, how much they look like girls/boys, how much they like to do the same thing as girls/boys, and how much they like to spend time with girls/boys.

Scores were calculated separately for similarity to their own gender ($\alpha = .75$) and similarity to the opposite gender ($\alpha = .75$). An average was taken across five questions for each gender, so if a child always said they were totally like girls, they would have an average score of a 5 for similarity to girls, and if they always said they were totally different from boys, they would have an average score of 1 for similarity to boys. For analyses examining inter-correlations between different measures, we also calculated a difference score for each participant, subtracting similarity to other gender from similarity to own gender, resulting in a scale from -4 to +4. This allowed us to recode the task onto a scale that was parallel to other tasks (i.e., identification with own gender vs. other gender). Participants who did not respond to at least three of five questions for each gender were excluded from the calculation of scores for that gender.

Gender Identity Implicit Association Test (IAT). We asked children to complete an Implicit Association Test (IAT) to determine the degree to which children implicitly associate themselves with girls versus boys (modeled from *1*). In the IAT, participants categorize pictures and words as quickly as they can, using two response keys on the computer. In the first block of trials, children sorted pictures of girls and boys using two different response keys. In the second block of trials, children were asked to sort words into the categories of “me” (“I,” “myself,” “my,” etc.) and “not me” (“them,” “theirs,” “other” etc.) using two response keys. In the third block of trials children used one button to respond to pictures of girls and “me” words, and another button to respond to pictures of boys and “not me” words (or the opposite pairing for half of participants). In the fourth block of trials, only pictures of boys and girls appeared, this time associated with the opposite keys from the first block. Finally, in the fifth block, children saw the opposite pairing from the third block—for example, now responding to pictures of boys and “me” words using one button and pictures of girls and “not me” words using another button.

For analyses, an IAT D score was computed using a standard algorithm (*11*). This score is based on the relative speed of responses in the third and fifth blocks (the cases in which gender and identity are paired). Higher positive scores indicated higher implicit identification with one’s own gender, and lower negative scores indicated higher implicit identification with the opposite gender. Scores at 0 indicated no clear association with one gender more than the other gender. Children who made errors on more than 30% of trials (n = 17), or who completed more than 10% of their responses in less than 300 ms (a speed that is too fast to process stimuli) (n = 3; 1 of whom is also included in those making more than 30% errors) were excluded. Because this test required participants to read, children under six years of age were not asked to complete this measure; children aged six and above who were not able to read were similarly not tested on this measure.

Additional measures

Rating of outfit at appointment. In addition to asking children to explicitly report on their clothing preferences, experimenters rated the clothing that children were wearing during the appointment. This rating was also on a 1 to 5 scale where (1) indicated the most stereotypical “boy” outfit (e.g., masculine sports attire, superhero costumes, or men’s formal wear), and where (5) indicated the most stereotypical “girl” outfit (e.g., princess costumes, frilly dresses and skirts). Beyond considering the type of outfit, researchers also considered color (e.g., pink), style (e.g., fitted vs. baggy), and accessories (e.g., sparkly headbands). Researchers could choose to use half points. For analyses purposed, boys’ scores were then reverse-scored so that for all participants higher scores reflected outfits stereotypical of the child’s own gender, and lower scores reflected outfits stereotypical of the child’s opposite gender.

For reliability purposes, we tried to have two researchers rate every child’s outfit. This happened for 562 cases, and there was high agreement, α = .97. Scores of coders were averaged to produce a final score. For 218 cases, there was only one outfit rating, and so the sole rater’s rating was used for analysis.

Parent measures

Demographics. Parents were asked to provide detailed demographic information including where their family currently lives (U.S. state or Canadian province), parental education level, family income, parental political orientation, child’s race (reported in Table S1), and whether or not the child was adopted. Family geographic locations were categorized into one of six categories (grouped by geography and to have roughly equal size groups in our sample): Northeast (CT, MA, MD, ME, NH, RI, VT, NJ, NY, PA, DC, DE, Ontario), Midwest/Upper Plains (IL, IN, MI, OH, WI, IA, KS, MN, MO, NE, ND, SD), Southeast (FL, GA, NC, SC, VA, WV, AL, KY, MS, TN, AR, LA, OK, TX), Mountain West (AZ, CO, ID, NM, MT, UT, NV, WY), Pacific NW (OR, WA, British Columbia, AK), Pacific South (CA, HI). For analyses, child race was re-coded as White (Non-Hispanic) and Non-White (the latter including multi-racial children for whom White is one race; since no single racial or ethnic group other than White included more than 10% of the sample). When two parents completed demographics questions, only one was used because parents typically agreed and there is no way to average responses given that race is a categorical variable. When possible, the mothers’ responses were utilized since the majority of parents who participated were mothers. If no mother was present, or if two mothers were present, the responses of the parent who was the primary study contact’s responses were utilized.

Time of social transition. Parents of transgender participants were asked at what age their child began to use pronouns corresponding to the gender opposite their assigned sex.

Parent report on child's gender identity now and in the future. Parents of all participants were asked whether they currently think of their child as a girl, a boy, or "other". If parents responded with 'other,' they were asked to specify in their own words. Parents were also asked what gender they thought their child would grow up to be and were given the same answer choices. Responses were scored with '1' point for responding with the child's identified gender, '-1' point for responding with the other gender, and '0' points for responding with 'other.' For participants who had two parents participating, parent scores were averaged, separately for the gender identity now and the gender identity in the future questions. Although this meant that transgender and sibling participants had an averaged score much more often than controls (because controls always had only one parent participating), having two control parents was not necessary since, for example, 99% of control parents responded with the expected gender for future identity (and 97% for current identity). In contrast, variation between parents' responses for transgender participants occurred occasionally (the equivalent numbers were 79% and 83% for parents of transgender participants).

Parent report of child's gendered behaviors and preferences. Parents were asked to answer eight questions about their child's preferences in various domains, and were asked to respond whether their child's preferences in each domain was girl-typed, boy-typed, or neutral. Specifically, parents were asked about their child's clothing preferences, peer preferences, toy preferences, media/TV/movie preferences, haircut style preferences, swimsuit preferences, avatars in video games and online, and favorite story protagonists. Parents were allowed to circle multiple response options, as they saw applicable. Responses were scored with '1' point if parents selected the child's gender, '-1' points if they selected the other gender, '0' points if they selected the neutral response or if they selected both genders. If parents circled one of the genders and the neutral option, their scores were averaged (i.e., they received either .5 or -.5, depending on the gender they selected). For participants who had two parents participating, parent scores were averaged; in cases where there were two parents, overall parent agreement was high, *r*'s > .405, *p's* < .001. All correlations are displayed in Table S3.

Table S3. Correlations Between Responses from Two Parents on Child's Gendered Behaviors and Preferences

Toy Preference	$r(340) = .574$, $p < .001$
Clothing Preference	$r(340) = .406$, $p < .001$
Peer Preference	$r(329) = .476$, $p < .001$

Gender Identity Questionnaire for Children (GIQC; adapted from (*12)*). Participants were asked 18 questions about their child's gender identity. These included questions about their child's behaviors, preferences, and expressed gender identity. Parents responded to each question on a 1 to 5 scale, rating either the frequency of their child's behaviors/preferences/gender identity expressions, or the frequency of which these behaviors/preferences/gender identity expressions were stereotypically aligned with boys, or stereotypically aligned with girls. In three of the questions, parents also had the option to respond with 'not applicable.' To make scoring consistent with the rest of the measures, responses were coded such that a score of '5' would indicate responses most aligned with the child's gender, and a score of '1' would indicate responses most aligned with the opposite gender (for transgender participants, a score of '1' indicated responses most aligned with the child's assigned sex). Of the 18 questions asked, 4 were excluded from the analyses, though are reported in the supplemental material. Two of these questions were excluded because they were also dropped by (*12*), the original paper reporting this scale, as they found that they did not load onto the same factor as other items. The other two questions were excluded because we added them as extra; specifically, Johnson et al. (*12*) gave different questionnaires depending on participant sex (i.e., assigned males' parents were asked "he states the wish to be a girl or a woman" and "he states that he is a girl or a woman"), whereas in the interest of convenience, all participants in our study received the same scale. Therefore, we asked all parents the two questions for assigned males and the two questions for assigned females resulting in an addition of two questions. However, for scoring purposes, we only counted the two questions that would have been originally asked based on a child's gender. As per Johnson et al. (*12*), on the three questions that allowed participants to respond with 'not applicable', if parents selected that option, those questions were not included in the calculation of the overall average across questions, and the denominator was adjusted accordingly. For participants with two parents participating in the study, parent scores were correlated, $r(197) = .652$, $p < .001$, and therefore their scores were averaged.

3. Comparing participant groups on child gender measures

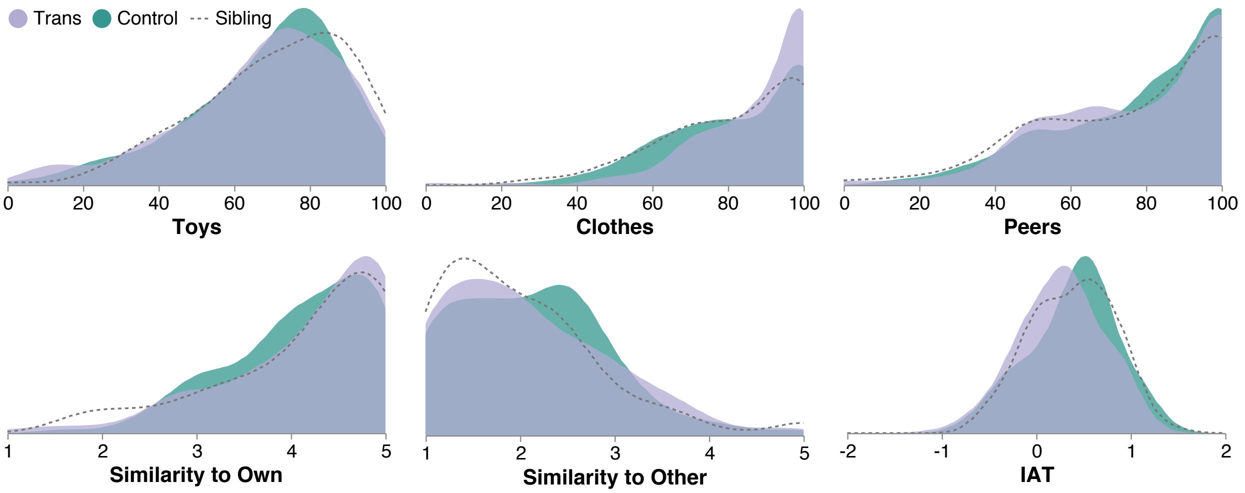

Fig S2. Density plots depicting transgender, control and sibling participants' scores on each of the measures. For all measures, greater numbers on the horizontal axis indicate greater identification and association with participants' own gender (*i.e.*, for transgender participants, their current gender), and lower numbers indicate greater identification and association with the "opposite" gender (*i.e.*, for transgender participants, the gender associated with their sex assigned at birth). As can be seen in the graphs, participants' scores in each group were largely overlapping across all measures. Note that this figure is a version of Fig. 1 (in the main paper) with boy participants' responses were reverse-coded. The scoring in this figure is consistent with how registered analyses in the main paper and supplement were conducted.

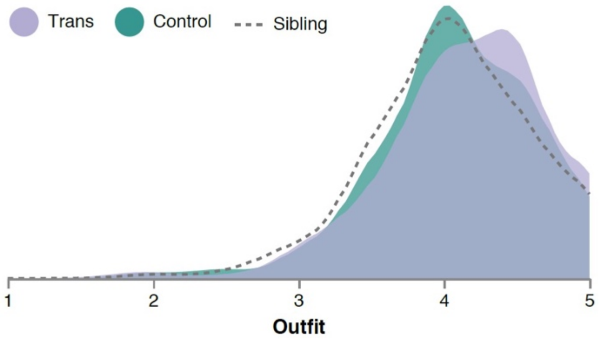

Fig S3. Outfit at appointment. The figure shows the frequency distribution of ratings of outfit at appointment by participant group. Raters scored each participant's outfit at appointment and scores were converted to a 1 (highly stereotypical of opposite gender) to 5 (highly stereotypical of own gender) scale. The distributions of ratings of outfit at appointment were very similar between participants groups.

4. Participants' responses on child measures compared to gender neutral

In the main paper, we reported in Table 1 participants' responses on each of the child measures compared to the midpoint of each task's scale. In Table S4, we provide detailed statistics associated with these analyses.

Table S4. Comparisons of participants' scores to chance or midpoints of the scale (meaning gender neutral, equally masculine/feminine) for each measure.

Task	Control	Transgender	Sibling
Toy preferences (Range 0-100)	$M = 68.42$, $SD = 20.18$ $t(273) = 15.11$, $p < .001$, d=.91	$M = 67.64$, $SD = 21.63$ $t(274) = 13.52$, $p < .001$, d=.82	$M = 70.92$, $SD = 19.94$ $t(162) = 13.39$, $p < .001$, d=1.05
Clothing preferences (Range 0-100)	$M = 82.74$, $SD = 17.67$ $t(273) = 30.67$, $p < .001$, d=1.85	$M = 87.97$, $SD = 15.43$ $t(274) = 40.81$, $p < .001$, d=2.46	$M = 81.63$, $SD = 18.36$ $t(162) = 22.00$, $p < .001$, d=1.72
Peer preferences (Range 0-100)	$M = 80.88$, $SD = 21.67$ $t(285) = 24.10$, $p < .001$, d=1.42	$M = 79.92$, $SD = 22.39$ $t(289) = 22.75$, $p < .001$, d=1.34	$M = 78.34$, $SD = 24.39$ $t(172) = 15.29$, $p < .001$, d=1.16
Similarity to own gender (Range 1 to 5)	$M = 4.11$, $SD = .75$ $t(269)$ =24.43 , p <.001, d=1.49	$M = 4.20$, $SD = .84$ $t(261) = 23.14$, p <.001, d=1.43	$M = 4.14$, $SD = .91$ $t(157)$ =15.74 , p <.001, d=1.25
Similarity to other gender (Range 1 to 5)	$M = 2.12$, $SD = .81$ $t(269) = -17.72$, $p < .001$, d=1.08	$M = 2.08$, $SD = .88$ $t(261) = -16.91$, $p < .001$, d=1.04	$M = 2.01$, $SD = .88$ $t(156) = -14.15$, $p < .001$, d=1.13
Implicit gender identity (Range ~ -2 to ~ +2)[1]	$M = .39$, $SD = .47$ $t(226) = 12.58$, $p < .001$, $d = .84$	$M = .26$, $SD = .45$ $t(207) = 8.40$, $p < .001$, d=.58	$M = .38$, $SD = .43$ $t(117) = 9.54$, $p < .001$, d=.88
Current gender identity[2]	83% $\chi^2(2)$=361.8, $p < .001$, $V = .76$	84% $\chi^2(2)$=369.6, $p < .001$, $V = .72$	87% $\chi^2(2)$=249.0, $p < .001$, $V = .81$
Future gender identity[2]	79% $\chi^2(2)$=313.4, $p < .001$, $V = .71$	80% $\chi^2(2)$=319.8, $p < .001$, $V = .71$	85% $\chi^2(2)$=226.4, $p < .001$, $V = .78$
Outfit at appointment (Range 1-5)	$M = 4.10$, $SD = .55$ $t(295) = 34.46$, $p < .001$, d=2.00	$M = 4.17$, $SD = .55$ $t(293) = 36.63$, $p < .001$, d=2.14	$M = 4.07$, $SD = .55$ $t(175) = 25.62$, $p < .001$, d=1.93

Notes: Higher scores on all measures indicate greater alignment with current gender identity, so in all cases these results indicate children saw themselves as more associated with their gender than chance responding and/or neutral responding would indicate.

[1] Technically, implicit gender identity scores could range above or below -/+ 2 however, in reality they seldom do.

[2] Because the current and future gender identity questions were categorical measures (participants could select one of the following three response options: boy, girl, other), these analyses compare the percentage of participants who responded with each possible response to a chance distribution (*i.e.*, a 33%, 33%, 33% distribution) with the use of Fischer's Exact Tests. Percentages in each cell indicate percentages of participants who responded with their own gender.

5. Child gender development examined by participant gender

Summary and take-home points. In this section, we describe findings from analyses examining gender differences. As per our registration, all analyses examining children's gender development included comparisons by participant group and participant gender. Due to space constraints, of all registered analyses, in the main paper we only included findings regarding participant group comparisons. Here, of the remaining registered analyses, we describe each measure with focus on findings relevant to participant gender. In our registration (https://osf.io/q2kuw/?view_only=c9f9df7d1e5a4f95ab893f28a81ce9e0), we did not expect to find any gender-based differences in children's gender development. However, as can be seen below, we found many gender differences, though the direction of these differences differed by measure. That is, on some measures, girl participants showed stronger association to their own gender, whereas on other measures, boys showed stronger association to their own gender. Additionally, these gender differences seldom interacted with participant group. Taken together, we believe that these findings suggest no consistent or interpretable gender difference.

Measures of gender identity

Explicit gender identity

Two Fisher's exact tests were used for the same reasons described above, to examine potential gender differences in participants' responses to the current and future gender identity questions. In contrast to our registered hypotheses, results showed that girls and boys differed in their responding when asked about their current gender identity ($p < .001$, V=.13) and when asked about their future gender identity ($p = .004$, V= .12). Tables S5 and S6 show percentages of participants who responded with each possible answer choice. Both groups showed a tendency to respond with their current gender for both the current (80% of girls; 90% of boys) and future (77% of girls, 86% of boys) identity questions. The difference between genders appears to be driven by the fact that on the current identity question 18% of girls and 9% of boys, and on the future identity question 21% of girls and 12% of boys replied with a response that was neither boy nor girl ("something else," e.g., "both boy and girl," "neither boy nor girl," "I don't know", "it changes,").

We conducted chi-square goodness of fit tests, separately for the current gender identity and future identity questions, to assess whether participants' responses (boy, girl, something else) differed from chance. Because there were gender differences in our previous analyses, as per our registration, we conducted separate analyses for girls and boys. Consistent with our registered hypotheses, both boys' (*p*s < .001) and girls' (*p*s < .001) responses differed from chance responding, meaning they were more likely to give some answers than others.

Table S5

Responses to Current Gender Identity Measure by Participant Group & Gender

Current Gender Identity		
	Boys	**Girls**
Cisgender Controls		
Boy	93.6%	1.9%
Girl	0.0%	77.7%
Neither	0.0%	1.5%
Both	2.8%	6.8%
It changes	0.9%	3.9%
I don't know	2.8%	8.3%
Transgender		
Boy	89.8%	1.5%
Girl	1.9%	81.0%
Neither	0.0%	1.0%
Both	4.6%	7.3%
It changes	0.0%	4.4%
I don't know	3.7%	4.9%
Cisgender Siblings		
Boy	87.9%	1.2%
Girl	0.9%	86.4%
Neither	0.0%	0.0%
Both	3.7%	6.2%
It changes	1.9%	2.5%
I don't know	5.6%	3.7%

Note. Participants in this study received one of two versions of this task. The first 409 participants received a version where they were asked "Are you a boy, girl, neither, both, it changes, or I don't know?" Because providing six answer choices at once seemed difficult for younger participants, we changed the measure so that the remaining 409 participants heard the questions "Are you a boy, girl, or something else?" In this version of the measure, if participants responded with "something else," they were then given the additional options of "neither," "both," "it changes," and "I don't know." In our registration, and for analysis purposes, the last four options ("Neither", "Both", "It changes", and "I don't know") were combined into one "something else" category for comparison to the "Boy" and "Girl" response categories.

Table S6

Responses to Future Gender Identity Measure by Participant Group & Gender

Future Gender Identity		
	Boys	**Girls**
Cisgender Controls		
Boy	88.9%	1.0%
Girl	0.9%	74.1%
Neither	0.0%	2.0%
Both	1.9%	5.4%
It changes	1.9%	3.4%
I don't know	6.5%	14.1%
Transgender		
Boy	85.0%	1.4%
Girl	0.1%	77.7%
Neither	0.0%	1.5%
Both	4.7%	4.9%
It changes	1.9%	2.5%
I don't know	7.5%	11.9%
Cisgender Siblings		
Boy	85.0%	1.2%
Girl	1.9%	84.0%
Neither	0.9%	1.2%
Both	4.7%	1.2%
It changes	0.9%	0.0%
I don't know	6.5%	12.3%

Note. For analysis purposes, the last four options ("Neither", "Both", "It changes", and "I don't know") were combined into one "Different Response" category for comparison to the "Boy" and "Girl" response categories.

Implicit gender identity

We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' Gender Identity IAT scores. In line with registered hypotheses, there were no gender differences, $F(1,547) = 2.19$, $p = .140$, $\eta_p{}^2 < .01$. We also did not find a significant participant group x participant gender interaction, $F(2,547) = 1.05$, $p = .352$, $\eta_p{}^2 < .01$.

Similarity to own- and other-gender children

Similarity to own gender. We conducted separate 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVAs on participants' perceived similarity to their own gender and the other gender. Contrary to our predictions, we found a significant main effect of participant gender on perceived similarity to own gender, $F(1,684) = 24.87$, $p < .001$, $\eta_p{}^2 = .04$, indicating that boys ($M = 4.33$, $SD = .72$) felt greater similarity to their own gender than girls did ($M = 4.04$, $SD = .87$). We did not find a significant participant group x gender interaction, $F(2,684) = 1.73$, $p = .179$, $\eta_p{}^2 < .01$.

Similarity to other gender. Our findings for similarity to other gender showed the same pattern of results. Counter to our registered hypotheses, we found a significant main effect of participant gender, $F(1,683) = 10.63$, $p = .001$, $\eta_p{}^2 = .02$, suggesting that boys ($M = 1.94$, $SD = .77$) perceived lower similarity to the other gender than girls did ($M = 2.17$, $SD = .89$). We found no significant participant group x gender interaction: $F(2,683) = 0.04$, $p = .958$, $\eta_p{}^2 < .01$.

Comparisons to the midpoint of scale. As per our registration, we calculated a difference score for each participant, subtracting perceived similarity to the other gender from perceived similarity to own gender to indicate whether children felt more similar to their own gender than the other gender. We conducted one-sample *t*-test comparisons of this difference score to the midpoint of the scale (0), separate for each gender, given the gender differences found in the ANOVAs reported above. In line with our registered hypotheses, we found that both girls, $t(420) = 28.79$, $p < .001$, $d = 1.40$, and boys, $t(267) = 33.75$, $p < .001$, $d = 2.39$, show greater perceived similarity to their own gender than the other gender.

Measures of gender-typed preferences

Toy preferences

We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' toy preferences. In contrast to our registered hypotheses, we found a significant main effect of participant gender, $F(1,706) = 62.75$, $p < .001$, $\eta_p{}^2 = .08$, indicating that boys' toy preferences ($M = 76.56$, $SD = 16.04$) were more stereotypically masculine than girls' preferences were stereotypically feminine ($M = 63.74$, $SD = 21.77$). There was not a significant interaction between participant group x participant gender, $F(2,706) = 1.22$, $p = .297$, $\eta_p{}^2 < .01$.

Given the significant gender differences, and the lack of a significant effect of participant group, we conducted a series of one-sample *t*-test comparisons to gender-neutral preferences (midpoint of scale = 50) for each gender group (collapsed across participant group). In line with our predictions, we found that both boys, $t(274)= 27.46$, $p < .001$, $d = 1.66$, and girls, $t(436) = 13.19$, $p < .001$, $d = 0.63$, showed preferences for toys stereotypically associated with their own gender. These findings were also consistent with the overall sample, $t(711) = 24.08$, $p < .001$, $d = 0.90$.

Clothing preferences

We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' clothing preferences. Contrary to our registered predictions, we found a significant main effect of participant gender, $F(1,706) = 12.05$, $p < .001$, $\eta_p{}^2 = .02$, indicating that boys' clothing preferences ($M = 81.28$, $SD = 16.25$) were less stereotypically masculine than girls' preferences were stereotypically feminine ($M = 86.54$, $SD = 17.50$). There was not a significant interaction of participant group x participant gender, $F(2,706) = 2.09$, $p = .124$, $\eta_p{}^2 < .01$.

Given the significant gender and participant group differences (the latter is reported in the main text), we conducted a series of one-sample *t*-test comparisons to gender neutral clothing preferences (midpoint of scale = 50) for each subgroup (transgender girls, transgender boys, control girls, control boys, sibling girls, sibling boys). In line with our predictions, we found that all groups were above the midpoint in their preferences for clothing stereotypically associated with their own gender: transgender girls, $t(183) = 31.54$, $p < .001$, $d = 2.33$; transgender boys, $t(90) = 27.59$, $p < .001$, $d = 2.89$; control girls, *t*(182) = 26.69, *p* < .001, *d* = 1.97; control boys, *t*(90) = 16.48, *p* < .001, *d* = 1.73; sibling girls, *t*(69) = 15.61, *p* < .001, *d* = 1.87; sibling boys, *t*(92) = 15.67, *p* < .001, *d* = 1.63.

Peer preferences

We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' peer preferences. In contrast to our registered predictions, we found a significant main effect of participant gender, *F*(1,743) = 10.26, *p* = .001, η_p^2 = .01, indicating that boys (*M* = 76.35, *SD* = 24.82) showed lower same-gender peer preferences compared to girls (*M* = 82.14, *SD* = 20.81). There was not a significant interaction of participant group x participant gender, *F*(2,743) = 0.03, *p* = .974, η_p^2 < .01.

Given the significant gender differences, and the lack of a significant effect of participant group, we conducted a series of one-sample *t*-test comparisons to chance (chance = 50) for each gender group (collapsed across participant group). We found that both boys, *t*(286) = 17.98, *p* < .001, *d* = 1.06, and girls, *t*(461) = 33.20, *p* < .001, *d* = 1.54, preferred same-gender peers, which was consistent with our predictions. We also found these same patterns of preferring same-gender peers in the overall sample, *t*(748) = 36.25, *p* < .001, *d* = 1.32.

Measures of gender-typed behavior

Outfit at appointment

We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on ratings of participants' outfits at appointments. We did find an unexpected significant main effect of participant gender, *F*(1,760) = 28.36, *p* < .001, η_p^2 = .03, indicating that boys' (*M* = 4.25, *SD* = .47) outfits were more stereotypically masculine than girls' outfits were stereotypically feminine (*M* = 4.03, *SD* = .58). Further, this main effect was qualified by a significant participant group x gender interaction, *F*(2,760) = 3.73, *p* = .002, η_p^2 = .01. Follow-up simple effects analyses showed that among transgender and control participants, boys' outfits at appointment (*M*s = 4.27 and 4.34, respectively) were rated as more stereotypically masculine than girls' outfits (*M*s = 4.10 and 3.97, respectively) were rated as stereotypically feminine (controls: *p* < .001; transgender: *p* = .020). However, there were no gender differences among siblings (boys, *M* = 4.12; girls, *M* = 3.99; *p* = .12).

Given the significant interaction effect, we conducted one-sample *t*-test comparisons to gender neutral clothing (midpoint of scale = 3) separately for boys and girls in each participant group. Results showed that girls and boys in all groups were more likely to wear outfits associated with their own gender, compared to gender neutral (transgender boys: *t*(101) = 28.77, *p* < .001, *d* = 2.85; transgender girls: *t*(191) = 26.28, *p* < .001, *d* = 1.90; control boys: *t*(101) = 31.10, *p* < .001, *d* = 3.08; control girls: *t*(193) = 24.19, *p* < .001, *d* = 1.74; sibling boys: *t*(97) = 22.44, *p* < .001, *d* = 2.27; sibling girls: *t*(77) = 14.36, *p* < .001, *d* = 1.63).

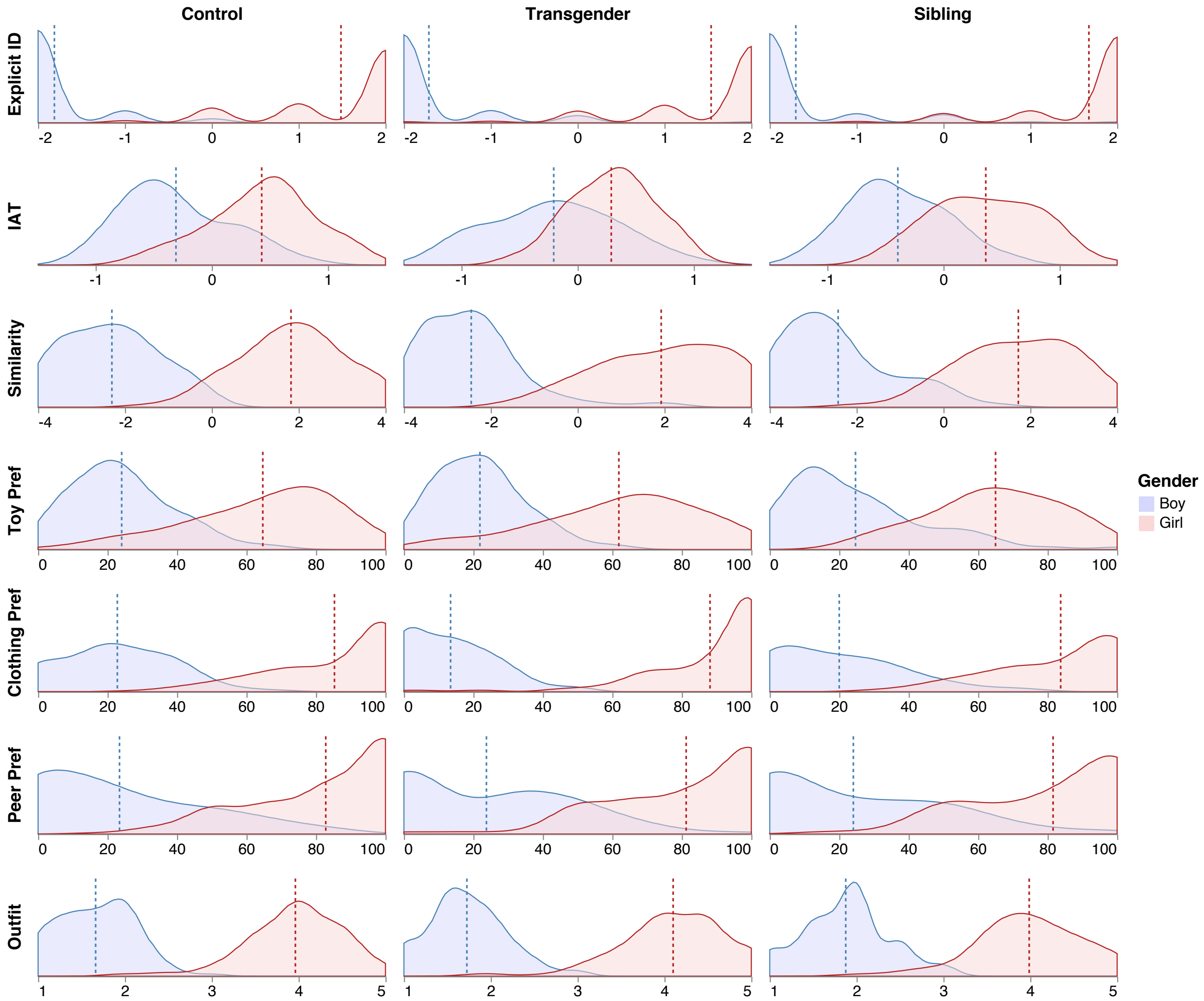

Fig. S4. Density plots depicting distributions of girls' and boys' responses on each child measure across the three participant groups. Higher scores on the x-axis indicate greater identification/association with girls, lower scores on the x-axis indicate greater identification/association with boys. Vertical dotted lines in graphs represent means.

6. Age-related changes in child measures

Summary and take-home points. Although we did not have any prior hypotheses regarding age-related changes in any of our measures, given that research has shown decreases in children's gender rigidity across development *(13-16)*, we examined developmental patterns in children's gender typing (Table S7 shows children's scores on each measure at each age; Table S8 shows correlation values). We found that participants' explicit identification with their own gender increased with age, though their implicit identification and their perceived similarity to their own vs. other did not change. This finding contrasts with previous research showing that girls' perceived similarity to their own gender decreases with age (while their similarity to the other gender increased with age), though is consistent with the finding that boys' perceived similarity does not change *(10)*.

In terms of gender-typed behaviors, we also found a mixture of age-related patterns. Whereas participants' preferences for gender-typed toys and peers did not change with age, their preferences for clothing and the ratings of their outfits at the time of appointment became less gender-typed as they grew older. Clothing preferences have been the focus of research, especially concerning girls' interest in pink frilly dresses, and this preference declines from ages 3 to 4, to ages 5 to 6, although boys show a later time of strong avoidance of feminine clothing *(17)*. The decline in children's preferences for gender-typed clothing is consistent with literature showing that children's endorsement of gender stereotypes declines with age *(15)*. However, although age-related declines in children's gender-typed preferences have been demonstrated by many researchers (*13-16*), some have also found that children's gender-typed preferences stay the same (*18*).

Table S7

Scores on Child Measures by Participant Group & Age

	Age (years)									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Toy Preference[1]										
Controls	76.56 (13.86) [4]	75.28 (18.35) [22]	76.71 (11.86) [39]	66.03 (18.85) [46]	71.60 (13.09) [46]	57.71 (23.08) [30]	69.92 (22.31) [32]	60.04 (25.62) [33]	70.31 (24.15) [20]	53.12 (22.10) [2]
Transgender	51.25 (12.02) [5]	57.89 (20.50) [19]	70.54 (18.94) [42]	66.46 (17.50) [41]	68.51 (19.84) [53]	61.42 (27.60) [29]	70.36 (23.27) [31]	73.11 (24.78) [33]	70.94 (20.81) [20]	65.62 (30.94) [2]
Siblings	82.03 (16.85) [8]	54.02 (21.60) [14]	64.58 (21.59) [21]	77.60 (11.45) [12]	71.80 (16.82) [28]	75.00 (18.54) [26]	73.21 (23.97) [21]	73.58 (17.03) [22]	70.54 (14.30) [7]	62.50 (32.68) [4]
Clothing Preference[1]										
Controls	76.56 (31.20) [4]	89.02 (16.84) [22]	90.71 (12.49) [39]	82.74 (19.42) [46]	83.56 (15.78) [46]	79.79 (18.33) [30]	81.84 (15.82) [32]	76.70 (18.51) [33]	77.19 (20.00) [20]	65.62 (4.42) [2]
Transgender	92.50 (13.55) [5]	90.46 (13.24) [19]	94.05 (10.60) [42]	91.46 (15.92) [41]	88.33 (15.41) [53]	77.80 (21.49) [29]	87.84 (11.75) [31]	84.66 (15.40) [33]	84.69 (13.67) [20]	81.25 (8.84) [2]
Siblings	84.38 (20.59) [8]	73.21 (28.42) [14]	81.25 (16.30) [21]	89.58 (12.31) [12]	85.94 (15.18) [28]	82.21 (17.29) [26]	84.82 (17.96) [21]	78.12 (17.01) [22]	76.79 (13.84) [7]	60.94 (26.21) [4]
Peer Preference[2]										
Controls	62.50 (34.36) [4]	81.75 (24.10) [21]	81.62 (24.42) [39]	79.22 (26.11) [47]	78.90 (19.49) [47]	78.92 (19.38) [34]	87.25 (18.83) [34]	79.28 (18.59) [37]	85.83 (18.16) [20]	83.33 (16.67) [3]
Transgender	83.33 (25.82) [6]	89.17 (18.16) [20]	82.17 (26.58) [43]	84.92 (21.40) [42]	82.20 (18.71) [50]	73.53 (24.66) [34]	78.92 (21.83) [34]	76.58 (19.82) [37]	68.25 (24.10) [21]	77.78 (9.62) [3]

Siblings	81.25 (25.88) [8]	67.86 (34.88) [14]	73.19 (25.49) [23]	75.64 (29.36) [13]	81.72 (25.22) [31]	78.22 (20.58) [30]	83.67 (22.86) [20]	82.61 (19.12) [23]	88.10 (12.60) [7]	54.17 (8.33) [4]
Similarity to Own[3]										
Controls	4.00 (1.04) [3]	4.08 (0.89) [19]	4.27 (0.64) [38]	4.10 (0.88) [46]	4.23 (0.70) [46]	4.07 (0.77) [30]	4.06 (0.67) [32]	3.84 (0.70) [34]	4.15 (0.71) [20]	4.70 (0.42) [2]
Transgender	2.33 (0.83) [3]	4.14 (0.63) [14]	4.21 (0.81) [38]	4.38 (0.91) [40]	4.33 (0.70) [52]	4.10 (1.05) [28]	4.16 (0.86) [32]	4.12 (0.80) [32]	4.11 (0.71) [21]	4.50 (0.14) [2]
Siblings	2.33 (0.61) [3]	3.96 (1.24) [13]	4.34 (0.77) [20]	4.55 (0.73) [12]	4.19 (0.95) [29]	4.06 (0.88) [26]	4.10 (0.92) [21]	4.17 (0.72) [23]	4.16 (1.07) [7]	4.20 (0.67) [4]
Similarity to Other[3]										
Controls	3.27 (0.99) [3]	2.48 (1.10) [19]	2.13 (0.98) [38]	2.10 (0.97) [46]	2.10 (0.72) [46]	2.18 (0.60) [30]	1.93 (0.57) [32]	2.07 (0.58) [34]	1.96 (0.82) [20]	2.40 (0.28) [2]
Transgender	2.93 (0.83) [3]	2.74 (1.10) [14]	2.02 (0.74) [38]	1.87 (0.93) [40]	2.05 (1.01) [52]	2.35 (0.98) [28]	2.05 (0.73) [32]	1.86 (0.59) [32]	2.14 (0.72) [21]	2.40 (0.57) [2]
Siblings	2.67 (1.03) [3]	2.49 (1.60) [13]	1.99 (1.12) [20]	1.95 (0.58) [12]	1.84 (0.53) [29]	2.19 (0.93) [26]	1.84 (0.78) [20]	1.97 (0.63) [23]	1.67 (0.63) [7]	1.95 (0.44) [4]
Appointment Outfit[4]										
Controls	4.44 (0.47) [4]	4.24 (0.37) [23]	4.26 (0.41) [39]	4.20 (0.43) [46]	4.05 (0.61) [46]	3.95 (0.52) [30]	3.96 (0.69) [32]	3.93 (0.45) [34]	4.16 (0.64) [20]	4.03 (0.72) [22]
Transgender	4.50 (0.47) [5]	4.33 (0.69) [19]	4.24 (0.63) [41]	4.30 (0.51) [42]	4.19 (0.45) [52]	4.03 (0.70) [29]	4.18 (0.44) [32]	4.05 (0.45) [33]	4.02 (0.51) [21]	3.99 (0.57) [20]
Siblings	4.41 (0.35) [8]	4.07 (0.49) [14]	4.16 (0.60) [20]	4.15 (0.63) [12]	4.03 (0.54) [29]	4.33 (0.44) [26]	3.96 (0.54) [21]	3.79 (0.48) [23]	4.06 (0.44) [8]	3.87 (0.71) [15]

Gender IAT[5]										
Controls	—	—	—	0.34 (0.49) [32]	0.43 (0.41) [44]	0.25 (0.51) [35]	0.43 (0.37) [33]	0.50 (0.52) [38]	0.34 (0.39) [22]	0.40 (0.54) [23]
Transgender	—	—	—	0.29 (0.35) [22]	0.14 (0.44) [48]	0.23 (0.46) [30]	0.33 (0.46) [30]	0.38 (0.41) [36]	0.45 (0.49) [21]	0.07 (0.44) [21]
Siblings	—	—	—	0.37 (0.28) [7]	0.44 (0.42) [24]	0.32 (0.40) [22]	0.47 (0.38) [19]	0.26 (0.51) [23]	0.48 (0.36) [8]	0.40 (0.53) [15]
Current Gender[6]										
Controls	75% [4]	91% [23]	80% [41]	74% [47]	81% [47]	71% [35]	89% [35]	89% [38]	91% [22]	96% [23]
Transgender	83% [6]	84% [19]	72% [43]	86% [42]	81% [54]	82% [33]	80% [35]	95% [37]	91% [23]	95% [21]
Siblings	86% [7]	86% [14]	79% [24]	77% [13]	91% [32]	80% [30]	86% [21]	96% [24]	100% [8]	100% [15]
Future Gender[6]										
Controls	50% [4]	87% [23]	72% [39]	66% [47]	72% [47]	80% [35]	86% [35]	87% [38]	86% [22]	100% [23]
Transgender	83% [6]	71% [17]	67% [43]	79% [42]	76% [54]	73% [33]	85% [34]	89% [36]	100% [23]	95% [21]
Siblings	86% [7]	71% [14]	75% [24]	77% [13]	88% [32]	77% [30]	95% [21]	92% [24]	100% [8]	93% [15]

Note. Means, standard deviations (in parentheses), and *n*s (in brackets) for Toy Preference, Clothing Preference, Peer Preference, Similarity to Own Gender, Similarity to Other Gender, and Appointment Outfit are presented for each participant group at each age. Percentage of participants providing their own gender in response to the Current and Future Gender Identity items are presented for each group at each age.

For some measures and some age groups, there are different numbers of cisgender control and transgender participants. This difference occurs because control participants were required to be within +/- 4 months of transgender participants' ages, which occasionally meant a pair had differing age in years. For example, a 4 year, 2-month-old transgender child could be paired with a 3 year, 11 month old control and the former would be listed as a 4-year-old while the latter would be listed as a 3-year-old in the tables. In all analyses age was recorded in months. Further, there was one 13-year-old cisgender control participant matched to a 12-year-old transgender participant and this 13-year-old is included in the set of 12-year-old cisgender controls in the table.

[1] Toy and Clothing Preference measures are scored on a 0-100 scale, such that 100 means a participant consistently selected items most associated with their own gender, and 0 means a participant consistently selected items most associated with the other gender.

[2] Peer Preference measure is scored as the percentage of time children selected own-gender targets.

[3] Similarity to Own Gender and Similarity to Other Gender measures are scored on a 1 (totally different) to 5 (totally the same) scale.
[4] Appointment Outfit measure is scored on a 1 (highly stereotypical of other gender) to 5 (highly stereotypical of own gender) scale.
[5] Gender IAT measure is scored such that higher positive scores indicate higher implicit identification with one's gender and lower negative scores will indicate higher implicit identification with the other gender, with scores around 0 indicating equal association with both genders.
[6] Percentages represent the percent of participants providing their own gender.

Table S8

Correlations between children's scores on each measure and their age

	Correlations with age
Explicit gender identity	**r(810) = .20, p < .001**
Implicit gender identity	*r*(553) = .03, *p* = .469
Similarity to own- and other-gender children	*r*(689) = .07, *p* = .084
Toy preferences[1]	*r*(712) = -.001, *p* = .978
Clothing preferences[1]	***r*(712) = -.18, *p* < .001**
Peer preferences	*r*(749) = -.03, *p* = .465
Outfit at appointment	***r*(766) = -.18, *p* < .001**

Note. Positive correlations mean older children showed more same-gender responding on that measure. Bold values show significant correlations.

[1] It is important to note that different sets of stimuli were used for younger and older children to make the toys and clothing age-appropriate, which might have influenced results.

Table S9. Child Measures Correlated with Time Since Transition for Transgender Participants

	Time Since Transition
Toy Preference	*r*(275) = -.06, *p* = .347
Clothing Preference	***r*(275) = -.12, *p* = .039**
Peer Preference	*r*(290) = -.05, *p* = .425
Similarity Composite[1]	*r*(262) = -.04, *p* = .499
Implicit Gender Identity	*r*(208) = .060, *p* =.390
Appointment Outfit	*r*(294) = -.04, *p* = .481
Explicit Identity Composite[2]	*r*(309) = .02, *p* = .737

Note. Partial correlations, controlling for participant age.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender, in line with our registered analyses.

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items.

7. Coherence among child measures for transgender participants and cisgender controls

Table S10. Correlation matrix of child measures for cisgender control participants (below the diagonal) and transgender participants (above the diagonal) coded in line with child gender

	1. Toy	**2. Clothing**	**3. Peer**	**4. Similarity**	**5. Implicit**	**6. Outfit**	**7. Explicit**
1. Toy Preference	-	***r*(275) = .30*****	***r*(269) = .17****	***r*(260) = .40*****	*r*(172) = .04	***r*(273) = .19***	*r*(269) = .09
2. Clothing Preference	***r*(274) = .49*****	-	***r*(269) = .23*****	***r*(260) = .31*****	*r*(172) = -.03	***r*(273) = .24*****	***r*(269) = .18***
3. Peer Preference	***r*(265) = .20****	***r*(265) = .27*****	-	***r*(254) = .29*****	*r*(184) = .11	*r*(268) = .07	***r*(282) = .19***
4. Similarity[1]	***r*(268) = .29*****	***r*(268) = .24*****	***r*(260) = .37*****	-	***r*(172) = .17***	***r*(260) = .12***	***r*(260) = .32*****
5. Implicit Gender Identity	*r*(189) = .01	*r*(189) = -.07	*r*(202) = .12	*r*(190) = -.09	-	*r*(191) = -.02	*r*(205) = .12
6. Appointment Outfit	***r*(274) = .39*****	***r*(274) = .20*****	*r*(266) = .02	***r*(270) = .26*****	***r*(210) = -.15***	-	***r*(288) = .21*****
7. Explicit Gender Identity[2]	***r*(272) = .21*****	***r*(272) = .12***	***r*(284) = .17****	***r*(268) = .42*****	*r*(227) = -.09	***r*(293) = .14***	-

**p* < .05, ***p* < .01, ****p* < .001

Note. Variables are coded such that higher values represent responses more in line with participants' own gender. Significant correlations are bolded.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender, in line with the registration.

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items, and ranges from -2 (responded with other gender twice) to 2 (responded with own gender twice) in line with the registration.

8. Alternate scoring of tasks for assessing coherence

Summary and take-home points. In our registered analyses, we coded each child measure in terms of participants' own gender versus the other gender. That is, in the analyses reported in the main paper, higher scores on any given task indicated that the child was showing strong identification or association with their own gender, and lower scores indicated the opposite. Past work has suggested that how one scores the IAT has considerable impact on its relation with other measures. As has been previously demonstrated by Rae and Olson (*4*), scoring Gender Identity IATs according to participants' own gender vs. other gender (as opposed to scoring it as boy vs. girl) reduces the amount of variability in the scores and reduces the magnitude of its correlation with other measures. Therefore, we decided to conduct additional exploratory analyses that were not registered, in which we used the coding system from previous research, to see whether our findings are affected. Below, we present these exploratory correlation results for each participant group separately (Tables S11-S12). Importantly, in these analyses, the measures were coded such that higher scores indicated identification or association with girls, and lower scores indicated identification or association with boys.

Our own results were consistent with the pattern observed by Rae and Olson (*4*). In our exploratory analyses, we also found that when IAT responses were scored as girls vs. boys, rather than own vs. other gender, the IAT significantly correlated with all other measures. Even more interestingly, this was the case for all measures, not just the IAT. That is, all correlations were much higher when measures were scored where lower scores were boy/masculine answers and higher scores were girl/feminine answers than when the scores ranged from "not my gender" to "my gender". The main reason this occurs is because of a statistical property of data like these where one has two distinct clusters of data—one for boys and one for girls. When coded from "not my gender" to "my gender" the two clusters substantially overlap but when coded as "boy" to "girl" the two clusters pull apart, creating a stronger correlation. We demonstrate this phenomenon in Fig S5.

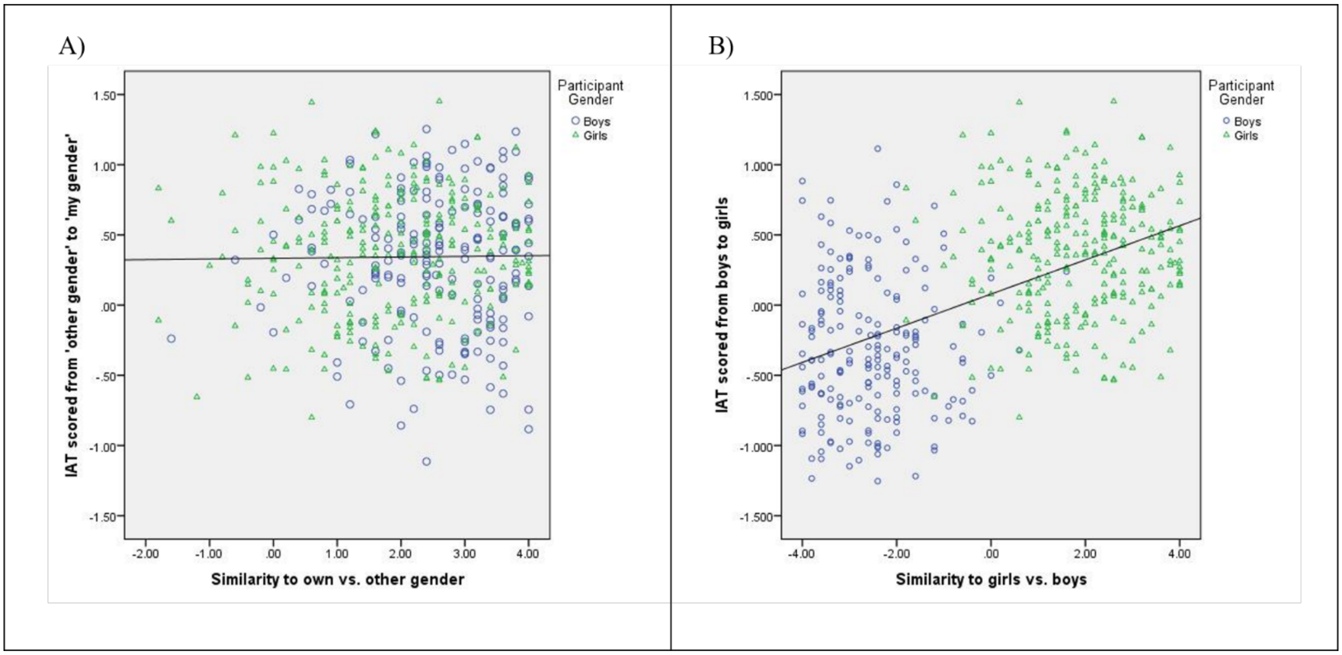

Fig S5. Correlations by different scoring methods

The figure shows the distribution of data points in correlations with similarity measures when the IAT is scored two different ways. Panel A shows the correlation between the IAT and similarity among all participants when coded according to the child's own gender: higher scores on the y-axis indicate implicit association with 'my gender' and lower scores on the y-axis indicate implicit association with the 'other gender'; higher scores on the x-axis indicate greater perceived similarity to own gender, and lower scores on the x-axis indicate greater perceived similarity to the other gender. Panel B shows the correlation for the two variables when coded from boy to girl: higher scores on the y-axis indicate implicit association with girls, whereas lower scores on the y-axis indicate implicit association with boys; higher scores on the x-axis indicate greater perceived similarity to girls, and lower scores on the x-axis indicate greater perceived similarity to boys.

We included these analyses because within the literature we have seen both approaches utilized (e.g., 19, 20) and we want to specifically call attention to this issue, echoing the comments from Rae and Olson (*4*) but extended to all measures of gender development. The field will need to unify behind one consistent strategy if one wants to be able to compare different papers to one another. We intentionally registered analyses using the more conservative analyses and make our final points focusing on this approach.

As can be seen in Tables S11 and S12, when measures were scored according to association with either gender, rather than with own vs. other gender, the correlation values between various gender cognition measures increased a great deal (e.g., for transgender participants, the correlation between toy and clothing preferences increased from *r* = .30 to *r* = .73.

Table S11. Correlation Matrix of Child Measures for Cisgender Control Participants (below the diagonal) and Transgender participants (above the diagonal) Coded in Line with Girl/Boy Associations

	1	2	3	4	5	6	7
1. Toy Preference	-	***r*(275) = .730 *p*<.001**	***r*(269) = .635 *p*<.001**	***r*(260) = .722 *p*<.001**	***r*(172) = .415 *p*<.001**	***r*(273) = .671 *p*<.001**	***r*(269) = .649 *p*<.001**
2. Clothing Preference	***r*(274) = .816 *p*<.001**	-	***r*(269) = .770 *p*<.001**	***r*(260) = .839 *p*<.001**	***r*(172) = .470 *p*<.001**	***r*(273) = .870 *p*<.001**	***r*(269) = .874 *p*<.001**
3. Peer Preference	***r*(265) = .667 *p*<.001**	***r*(265) = .757 *p*<.001**	-	***r*(254) = .761 *p*<.001**	***r*(184) = .473 *p*<.001**	***r*(268) = .724 *p*<.001**	***r*(282) = .756 *p*<.001**
4. Similarity[1]	***r*(268) = .692 *p*<.001**	***r*(268) = .814 *p*<.001**	***r*(260) = .807 *p*<.001**	-	***r*(172) = .532 *p*<.001**	***r*(260) = .779 *p*<.001**	***r*(260) = .835 *p*<.001**
5. Gender Identity IAT	***r*(189) = .449 *p*<.001**	***r*(189) = .485 *p*<.001**	***r*(202) = .555 *p*<.001**	***r*(190) = .512 *p*<.001**	-	***r*(191) = .448 *p*<.001**	***r*(205) = .494 *p*<.001**
6. Appointment Outfit	***r*(274) = .737 *p*<.001**	***r*(274) = .843 *p*<.001**	***r*(266) = .733 *p*<.001**	***r*(270) = .822 *p*<.001**	***r*(210) = .508 *p*<.001**	-	***r*(288) = .863 *p*<.001**
7. Explicit Gender Identity[2]	***r*(272) = .684 *p*<.001**	***r*(272) = .818 *p*<.001**	***r*(284) = .764 *p*<.001**	***r*(268) = .862 *p*<.001**	***r*(227) = .552 *p*<.001**	***r*(293) = .839 *p*<.001**	-

Note. Variables are coded such that higher values represent responses more in line with girls. That is, higher numbers represent choosing more girly toys and clothes, preferring girls as friends, feeling more similar to girls, wearing a more girly outfit to the appointment, and more often identifying as a girl. All correlations were significant.

[1] The similarity variable used here is the difference between similarity to girls and similarity to boys (similarity to girls minus similarity to boys).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with “boy” twice) to 2 (responded with “girl” twice).

Table S12. Correlation Matrix of Child Measures for Cisgender Sibling Participants Coded in Line with Girl/Boy Associations

	Cisgender Sibling Participants						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1. Toy Preference	-						
2. Clothing Preference	*r*(163) = .791 *p*<.001	-					
3. Peer Preference	*r*(160) = .737 *p*<.001	*r*(160) = .781 *p*<.001	-				
4. Similarity[1]	*r*(154) = .729 *p*<.001	*r*(154) = .822 *p*<.001	*r*(152) = .781 *p*<.001	-			
5. Gender Identity IAT	*r*(96) = .533 *p*<.001	*r*(96) = .551 *p*<.001	*r*(103) = .574 *p*<.001	*r*(97) = .537 *p*<.001	-		
6. Appointment Outfit	*r*(162) = .703 *p*<.001	*r*(162) = .848 *p*<.001	*r*(160) = .743 *p*<.001	*r*(156) = .774 *p*<.001	*r*(110) = .579 *p*<.001	-	
7. Explicit Gender Identity[2]	*r*(162) = .717 *p*<.001	*r*(162) = .821 *p*<.001	*r*(171) = .776 *p*<.001	*r*(157) = .840 *p*<.001	*r*(118) = .666 *p*<.001	*r*(175) = .838 *p*<.001	-

Note. Variables are coded such that higher values represent responses more in line with girls. That is, higher numbers represent choosing more girly toys and clothes, preferring girls as friends, feeling more similar to girls, wearing a more girly outfit to the appointment, and more often identifying as a girl. All correlations are significant.

[1] The similarity variable used here is the difference between similarity to girls and similarity to boys (similarity to girls minus similarity to boys).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with “boy” twice) to 2 (responded with “girl” twice)

9. Demographic comparisons

Summary and take-home points. Few studies of gender development have had the power or diversity within a sample to examine whether gender typing varies as a function of demographic variables (e.g., race, geographic location, etc.). Although there may be little reason to believe that children from higher or lower income backgrounds, for example, would identify any more or less strongly with their gender, studies have occasionally found some demographic differences in gender development (e.g., *17, 21-23*). The current study provides both a detailed description of demographic characteristics of our sample of transgender and cisgender children and an examination of how variations in demographic variables (i.e., race, location, parental education level and political ideology, and household income) might relate to variations in children's gender typing (i.e., gender identification and gender-typed preferences/behaviors). Due to the dearth of prior evidence, the analyses presented in this chapter are exploratory. That is, although the analyses were registered https://osf.io/q2kuw/?view_only=c9f9df7d1e5a4f95ab893f28a81ce9e0), there were no registered hypotheses for any of the findings.

The overall take-home from these findings is just how remarkably consistent they are. Regardless of whether participants are transgender or cisgender, whether they are White or non-White, where they live, what their family income is, or what their parent education levels are, children showed very similar levels of gender development across a variety of measures. When differences were observed, the effects tended to be small. Thus, these results complement those in the main paper, showing just how robust the present findings are across demographic factors, age, and distinctions between transgender, sibling, and control participants.

Method

Parents were asked to provide detailed demographic information including where their family currently lives (U.S. state or Canadian province), parental education level, family income, parental political orientation, child's race (reported in Table S1, and whether or not child was adopted. Family geographic locations were categorized into one of six categories (grouped by geography and to have roughly equal size groups in our sample): Northeast (CT, MA, MD, ME, NH, RI, VT, NJ, NY, PA, DC, DE, Ontario), Midwest/Upper Plains (IL, IN, MI, OH, WI, IA, KS, MN, MO, NE, ND, SD), Southeast (FL, GA, NC, SC, VA, WV, AL, KY, MS, TN, AR, LA, OK, TX), Mountain West (AZ, CO, ID, NM, MT, UT, NV, WY), Pacific NW (OR, WA, British Columbia, AK), Pacific South (CA, HI). For analyses, child race was re-coded as White (Non-Hispanic) and Non-White (the latter including multi-racial children for whom White is one race; since no single racial or ethnic group other than White included more than 10% of the sample). When two parents completed demographics questions, only one was used because parents typically agreed and there is no way to average responses given that race is a categorical variable. When possible, the mothers' responses were utilized since the majority of parents who participated were mothers. If no mother was present, or if two mothers were present, the responses of the parent who was the primary study contact's responses were utilized.

Results

Relations between demographics and child measures

Registered Analysis plan

For each measure, we conducted an identical set of analyses. First, to compare White and non-White participants' scores on a measure, we conducted independent-samples *t*-tests. Second, we conducted one-way analyses of variance to examine potential differences based on geographic location. Finally, we conducted 3 sets of correlational analyses between each measure and parent education level, parent political orientation, and household income. Analyses were conducted separately for each participant group. For comparisons by geographic location, however, we did not include control participants in the analyses as they were all recruited within the same geographic location. For each measure, detailed statistics are presented in the specified tables.

Explicit gender identity

Because demographic differences were not registered as a primary aim of the current work, were largely exploratory, and due to the large number of demographic analyses, we registered a conservative alpha threshold of $p < .005$. We found no significant differences at this threshold in participants' scores as a function of demographics for any of the participant groups (see Table S13 for details).

Table S13. Scores on Explicit Identity Composite by Participant Group and Demographics

	Explicit Identity Composite					
	Mean (SD)			Statistics		
	Controls	**Transgender**	**Siblings**	**Controls**	**Transgender**	**Siblings**
Child's Race[1]				$t(161.3) = 0.04$ $p = .971$, $d <$ 0.01	$t(156.4) = 0.86$ $p = .390$, $d =$ 0.11	$t(129.2) = 1.14$ $p = .257$, $d = 0.16$
White	1.61 (0.71) [218]	1.63 (0.73) [211]	1.66 (0.75) [130]			
Non-White	1.60 (0.77) [93]	1.54 (0.87) [95]	1.78 (0.57) [54]			
Geographic Location[2]				—	$F(5,301) = 1.96$ $p = .084$, $\eta_p{}^2 =$ .03	$F(5,179) = .489$ $p = .784$, $\eta_p{}^2 = .01$
Northeast	—	1.50 (0.86) [46]	1.68 (0.63) [25]	—		
Midwest/Upper Plains	—	1.76 (0.56) [62]	1.80 (0.50) [45]			
Southeast	—	1.65 (0.76) [48]	1.72 (0.84) [29]			
Mountain West	—	1.74 (0.55) [39]	1.61 (0.74) [28]			
Pacific Northwest	1.60 (0.72) [311]	1.56 (0.72) [62]	1.57 (0.88) [35]			
Pacific South	—	1.36 (1.08) [50]	1.70 (0.70) [23]			
Parent Political Ideology[3]				$r(309) = -.05$ $p = .422$	$r(308) = .15$ $p = .011$	$r(187) = .100$ $p = .170$
Liberal	1.60 (0.71) [196]	1.55 (0.81) [266]	1.68 (0.73) [158]			
Moderate	1.63 (0.74) [103]	1.88 (0.46) [41]	1.70 (0.61) [27]			
Conservative	1.30 (0.82) [10]	*	*			
Parent Education Level[4]				$r(117) = .10$ $p = .279$	$r(117) = -.05$ $p = .623$	$r(68) = -.04$ $p = .758$
Some schooling	—	—	—			
High school diploma	*	*	*			
Some college/Associate's degree	1.71	1.71	1.71			

	(0.69) [17]	(0.56) [21]	(0.73) [14]			
College/Bachelor’s degree	1.80	1.63	1.92			
	(0.52) [56]	(0.88) [38]	(0.41) [24]			
Advanced degree (MA, MD, PhD, etc.)	1.86	1.63	1.72			
	(0.41) [44]	(0.70) [57]	(0.75) [29]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(311) = .06	*r*(307) = -.06	*r*(185) = -.14
Less than $25,000	1.67	1.73	1.78	*p* = .276	*p* = .314	*p* = .061
	(0.82) [6]	(0.65) [11]	(0.44) [9]			
$25,001 to $50,000	1.53	1.65	1.78			
	(0.83) [15]	(0.66) [31]	(0.55) [18]			
$50,001 to $75,000	1.46	1.55	1.91			
	(0.79) [39]	(0.85) [64]	(0.38) [33]			
$75,001 to $125,000	1.63	1.72	1.66			
	(0.72) [98]	(0.68) [95]	(0.76) [58]			
More than $125,000	1.64	1.49	1.57			
	(0.69) [151]	(0.85) [106]	(0.84) [67]			

[*] Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] Two control parents circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). These data points are excluded from the table, but were scored as “3.5” on the income scale for analyses purposes.

Implicit gender identity

As can be seen in Table S14, we found no differences in participants' IAT scores based on race, geographic location, parent education level, parent political orientation, or household income for control, transgender, or sibling groups.

Table S14. Scores on IAT Measure by Participant Group and Demographics

	IAT					
	Mean (SD)			Statistics		
	Controls	**Transgender**	**Siblings**	**Controls**	**Transgender**	**Siblings**
Child’s Race[1]				$t(107.5) = 0.20$	$t(133.1) = 0.15$	$t(49.81) = 0.28$
White	.40 (0.45) [160]	.26 (0.47) [143]	.39 (0.44) [87]	$p = .844$, $d = 0.03$	$p = .878$, $d = 0.02$	$p = .782$, $d = 0.06$
Non-White	.38 (0.51) [66]	.25 (0.41) [62]	.36 (0.42) [29]			
Geographic Location[2]				—	$F(5,200) = 1.48$	$F(5,110) = 1.22$
Northeast	—	.38 (0.48) [27]	.21 (0.43) [13]	—	$p = .199$, $\eta_p{}^2 = .04$	$p = .307$, $\eta_p{}^2 = .05$
Midwest/Upper Plains	—	.34 (0.42) [44]	.28 (0.48) [31]			
Southeast	—	.14 (0.42) [33]	.42 (0.36) [17]			
Mountain West	—	.33 (0.43) [28]	.44 (0.50) [22]			
Pacific Northwest	.39 (0.47) [226]	.21 (0.49) [45]	.46 (0.43) [19]			
Pacific South	—	.18 (0.43) [29]	.51 (0.29) [14]			
Parent Political Ideology[3]				$r(225) = .06$	$r(207) = -.02$	$r(117) = -.03$
Liberal	.37 (0.49) [138]	.27 (0.46) [181]	.39 (0.44)[100]	$p = .337$	$p = .751$	$p = .767$
Moderate	.45 (0.44) [79]	.17 (0.34) [26]	.26 (0.42) [15]			
Conservative	.29 (0.40) [8]	*	*			
Parent Education Level[4]				$r(79) = .02$	$r(70) = -.13$	$r(37) = -.18$
Some schooling	—	—	—	$p = .851$	$p = .276$	$p = .283$
High school diploma	*	*	*			
Some college/Associate’s degree	.28	.29	.51			

	(0.38) [11]	(0.46) [10]	(0.39) [10]			
College/Bachelor’s degree	.36	.30	.46			
	(0.57) [39]	(0.49) [25]	(0.60) [13]			
Advanced degree (MA, MD, PhD, etc.)	.34	.16	.21			
	(0.44) [29]	(0.57) [34]	(0.45) [13]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(226) = .06	*r*(206) = .11	*r*(116) = .004
Less than $25,000	*	.01	.15	*p* = .333	*p* = .122	*p* = .964
		(0.50) [6]	(0.40) [5]			
$25,001 to $50,000	.20	.15	.47			
	(0.38) [13]	(0.53) [19]	(0.53) [9]			
$50,001 to $75,000	.36	.24	.46			
	(0.44) [28]	(0.40) [43]	(0.46) [23]			
$75,001 to $125,000	.41	.28	.34			
	(0.46) [69]	(0.41) [64]	(0.43) [36]			
More than $125,000	.41	.30	.38			
	(0.58) [111]	(0.48) [74]	(0.42) [43]			

* Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] Two control parents circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). These data points are excluded from the table, but were scored as “3.5” on the income scale for analyses purposes.

Similarity to own- and other-gender children

For all three groups, we found no differences at the registered $p < .005$ threshold in similarity to own vs. other gender based on race, geographic location, parent education level, political ideology, or household income (see Table S15).

Table S15. Scores on Similarity Composite by Participant Group and Demographics

	Similarity Composite					
	Mean (SD)			**Statistics**		
	Controls	**Transgender**	**Siblings**	**Controls**	**Transgender**	**Siblings**
Child's Race[1]				*t*(141.9) = 0.69 *p* = .489, *d* = .09	*t*(153.3) = 0.69 *p* = .489. *d* = .09	*t*(86.25) = 0.77 *p* = .446, *d* = 0.13
White	2.03 (1.17) [190]	2.07 (1.39) [179]	2.20 (1.32) [108]			
Non-White	1.92 (1.20) [79]	2.20 (1.41) [81]	2.02 (1.34) [47]			
Geographic Location[2]				—	*F*(5,255) = 1.01 *p* = .414, η_p^2 = .02	*F*(5,150) = 1.31 *p* = .261, η_p^2 = .04
Northeast	—	2.19 (1.52) [43]	1.60 (1.57) [24]	—		
Midwest/Upper Plains	—	2.28 (1.28) [57]	2.23 (1.20) [41]			
Southeast	—	2.10 (1.53) [42]	1.96 (1.02) [24]			
Mountain West	—	2.23 (1.27) [35]	2.10 (1.47) [22]			
Pacific Northwest	2.00 (1.18) [269]	2.06 (1.34) [46]	2.43 (1.22) [26]			
Pacific South	—	1.68 (1.42) [38]	2.38 (1.44) [19]			
Parent Political Ideology[3]				*r*(267) = -.003	*r*(260) = .15, *p* = .014	*r*(156) = .09
Liberal	1.98 (1.19) [176]	2.04 (1.37) [230]	2.07 (1.35) [135]	*p* = .956		*p* = .286
Moderate	2.05 (1.20) [82]	2.56 (1.51) [30]	2.53 (1.00) [19]			
Conservative	1.73 (0.82) [9]	*	*			
Parent Education Level[4]				*r*(110) = -.04	*r*(106) = -.14	*r*(58) = -.11
Some schooling	—	—	—	*p* = .686	*p* = .147	*p* = .413
High school diploma	*	*	*			
Some college/Associate's degree	2.13	2.13	2.07			

	(0.82) [17]	(1.44) [18]	(1.22) [12]			
College/Bachelor’s degree	1.94	2.08	2.51			
	(1.18) [51]	(1.46) [37]	(1.61) [17]			
Advanced degree (MA, MD, PhD, etc.)	1.96	1.76	1.88			
	(1.13) [42]	(1.47) [50]	(1.50) [28]			
Other	—	—	—			
More than one	—	—	—			
Household Annual Income[5]				*r*(269) = .09	*r*(261) = .03	*r*(156) = -.04
Less than $25,000	2.10	2.73	2.33	*p* = .147	*p* = .598	*p* = .633
	(1.24) [6]	(1.22) [9]	(1.37) [8]			
$25,001 to $50,000	1.61	1.90	2.57			
	(1.54) [11]	(1.41) [25]	(1.10) [14]			
$50,001 to $75,000	1.65	1.97	1.88			
	(0.98) [33]	(1.33) [52]	(1.41) [29]			
$75,001 to $125,000	2.08	2.02	2.10			
	(1.23) [86]	(1.53) [79]	(1.25) [48]			
More than $125,000	2.07	2.24	2.14			
	(1.15) [132]	(1.32) [96]	(1.38) [57]			

[*] Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] Two control parents circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). These data points are excluded from the table, but were scored as “3.5” on the income scale for analyses purposes.

Toy preferences

Participants' toy preferences did not vary as a function of race, geographic location, parent education, or household income in any of the three participant groups (see Table S16). However, there was a small, significant positive correlation between parent political ideology and toy preferences for transgender participants, *r*(274) = .17, *p* = .004. This effect indicated that transgender participants whose parents were less liberal showed more gender-typed preferences in toys. The correlation for transgender participants between toy preferences and political ideology was still significant and positive when the one conservative participant was excluded, *r*(273) = .18, *p* = .003. For cisgender controls and siblings, there was not a significant correlation between parent political ideology and toy preferences.

Table S16. Scores on Toy Preference Measure by Participant Group and Demographics

	Toy Preferences					
	Mean (SD)			Statistics		
	Controls	**Transgender**	**Siblings**	**Controls**	**Transgender**	**Siblings**
Child's Race[1]				$t(170.7) = 0.94$ $p = .347$, $d = 0.12$	$t(144.4) = 0.22$ $p = .822$, $d = 0.03$	$t(117.4) = 0.33$ $p = .740$, $d = 0.05$
White	69.13 (20.77) [190]	67.03 (21.23) [189]	70.48 (21.44) [112]			
Non-White	66.72 (18.90) [83]	67.71 (23.74) [84]	71.51 (16.50) [49]			
Geographic Location[2]				—	$F(5,268) = 0.51$	$F(5,156) = 1.09$
Northeast	—	69.46 (23.93) [44]	66.93 (20.31) [24]	—	$p = .766$, $\eta_p^2 = .01$	$p = .369$, $\eta_p^2 = .03$
Midwest/Upper Plains	—	65.78 (21.70) [61]	70.63 (18.96) [42]			
Southeast	—	68.31 (21.58) [43]	67.07 (22.19) [26]			
Mountain West	—	70.27 (18.89) [37]	78.39 (13.54) [24]			
Pacific Northwest	68.40 (20.22) [273]	66.93 (22.99) [48]	71.88 (24.74) [28]			
Pacific South	—	63.72 (22.50) [41]	69.79 (16.22) [18]			
Parent Political Ideology[3]				$r(271) = -.01$	$r(274) = .17$, $p = .004$	$r(162) = -.02$
Liberal	68.38 (18.93) [176]	66.42 (22.13) [239]	71.27 (19.65) [138]	$p = .876$		$p = .771$
Moderate	68.46 (22.97) [86]	73.16 (20.39) [34]	67.05 (21.92) [22]			
Conservative	68.06 (20.36) [9]	*	*			
Parent Education Level[4]				$r(110) = -.05$	$r(110) = .02$	$r(61) = -.25$
Some schooling	—	—	—	$p = .606$	$p = .812$	$p = .052$
High school diploma	*	*	*			
Some college/Associate's degree	71.69	66.78	74.48			

	(17.69) [17]	(16.28) [19]	(21.89) [12]			
College/Bachelor's degree	67.52	64.80	71.25			
	(20.43) [51]	(24.76) [38]	(17.13) [20]			
Advanced degree (MA, MD, PhD, etc.)	68.01	67.43	62.50			
	(17.47) [42]	(20.01) [52]	(20.27) [28]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(273) = .09	*r*(274) = .01	*r*(162) = -.14
Less than $25,000	68.75	80.00	80.56	*p* = .129	*p* = .824	*p* = .067
	(20.54) [6]	(11.33) [10]	(15.13) [9]			
$25,001 to $50,000	59.66	68.97	70.54			
	(25.82) [11]	(22.06) [29]	(16.88) [14]			
$50,001 to $75,000	67.59	63.10	72.46			
	(21.57) [34]	(22.83) [52]	(19.82) [32]			
$75,001 to $125,000	66.72	63.41	72.70			
	(21.85) [86]	(23.35) [82]	(19.71) [49]			
More than $125,000	70.42	70.79	66.77			
	(18.28) [135]	(20.25) [101]	(21.14) [58]			

* Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child's race. Welch's *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that "Liberal" consists of those responding 1-2, "Moderate" consists of those responding 3-5, and "Conservative" consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding "Other" or "More than one" to the parent education item.

[5] One control parent circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). This data point is excluded from the table, but was scored as "3.5" on the income scale for analyses purposes.

Clothing preferences

Across all three groups, participants' clothing preferences did not differ by race, geographic location, parent education level, parent political ideology, or household income (see Table S17 for detailed statistics).

Table S17. Scores on Clothing Preference Measure by Participant Group and Demographics

	Clothing Preferences					
	Mean (SD)			**Statistics**		
	Controls	**Transgender**	**Siblings**	**Controls**	**Transgender**	**Siblings**
Child's Race[1]				*t*(164) = 0.10	*t*(148.1) = 0.73	*t*(93.76) = 0.64
White	82.63	88.28	80.92	*p* = .922,	*p* = .464,	*p* = .526,
	(17.99) [190]	(15.68) [189]	(18.59) [112]	*d* = 0.01	*d* = .10	*d* = 0.11
Non-White	82.86	86.68	82.91			
	(17.09) [83]	(17.02) [84]	(18.12) [49]			
Geographic Location[2]				—	*F*(5,268) = 0.48	*F*(5,156) = 0.94
Northeast	—	86.93	82.03	—	*p* = .788,	*p* = .454,
		(19.24) [44]	(19.44) [24]		η_p^2 = .01	η_p^2 = .03
Midwest/Upper Plains	—	89.65	77.08			
		(13.59) [61]	(21.37) [42]			
Southeast	—	88.23	81.01			
		(16.88) [43]	(16.91) [26]			
Mountain West	—	88.01	85.42			
		(12.18) [37]	(15.05) [24]			
Pacific Northwest	82.70	87.50	82.37			
	(17.69) [273]	(19.55) [48]	(19.77) [28]			
Pacific South	—	84.71	86.11			
		(14.92) [41]	(12.42) [18]			
Parent Political Ideology[3]				*r*(271) = .08	*r*(274) = .07	*r*(162) = .08
Liberal	82.19	87.49	81.43	*p* = .185	*p* = .270	*p* = .328
	(17.38) [176]	(15.80) [239]	(18.67) [138]			
Moderate	82.56	88.60	80.97			
	(18.86) [86]	(19.37) [34]	(17.09) [22]			
Conservative	92.36	*	*			
	(10.26) [9]					
Parent Education Level[4]				*r*(110) = -.03	*r*(110) = .07	*r*(61) = -.10
Some schooling	—	—	—	*p* = .754	*p* = .470	*p* = .428
High school diploma	*	*	*			
Some college/Associate's degree	84.19	88.49	81.77			
	(16.99) [17]	(14.62) [19]	(18.36) [12]			

College/Bachelor’s degree	80.39	86.13	91.88			
	(18.63) [51]	(18.27) [38]	(13.77) [20]			
Advanced degree (MA, MD, PhD, etc.)	81.60	90.14	80.58			
	(16.43) [42]	(12.82) [52]	(16.78) [28]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(273) = .08	*r*(274) = .06	*r*(162) = -.11
Less than $25,000	79.17	90.62	86.81	*p* = .177	*p* = .304	*p* = .177
	(23.27) [6]	(11.88) [10]	(17.80) [9]			
$25,001 to $50,000	69.89	85.13	80.80			
	(22.85) [11]	(16.65) [29]	(15.97) [14]			
$50,001 to $75,000	83.27	84.70	84.77			
	(16.76) [34]	(18.29) [52]	(15.79) [32]			
$75,001 to $125,000	83.75	88.80	81.89			
	(17.25) [86]	(16.04) [82]	(18.19) [49]			
More than $125,000	83.10	88.68	78.99			
	(17.41) [135]	(15.48) [101]	(20.57) [58]			

[*] Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] One control parent circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). This data point is excluded from the table, but was scored as “3.5” on the income scale for analyses purposes.

Peer preferences

We did not find any significant differences at the $p < .005$ threshold in participants' peer preferences based on any of the demographics we measured (see Table S18 for all statistics)..

Table S18. Scores on Peer Preference Measure by Participant Group and Demographics

	Peer Preferences					
	Mean (SD)			Statistics		
	Controls	Transgender	Siblings	Controls	Transgender	Siblings
Child's Race[1]				*t*(160.6) = 0.62 *p* = .534, *d* = 0.08	*t*(183.2) = 0.18 *p* = .861, *d* = 0.02	*t*(101.2) = 0.24 *p* = .810, *d* = 0.04
White	80.38 (22.18) [203]	79.39 (23.14) [196]	78.73 (24.64) [118]			
Non-White	82.09 (20.58) [82]	79.89 (22.45) [92]	77.78 (23.01) [51]			
Geographic Location[2]				—	*F*(5,283) = 0.86 *p* = .506, η_p^2 = .02	*F*(5,164) = 0.19 *p* = .966, η_p^2 = .01
Northeast	—	80.68 (26.15) [44]	75.33 (25.06) [25]	—		
Midwest/Upper Plains	—	83.06 (21.41) [61]	78.17 (25.63) [42]			
Southeast	—	79.84 (21.07) [43]	78.40 (21.09) [27]			
Mountain West	—	78.24 (25.76) [36]	77.78 (21.80) [24]			
Pacific Northwest	80.87 (21.71) [285]	74.70 (21.55) [56]	80.94 (26.79) [32]			
Pacific South	—	80.82 (22.42) [49]	80.83 (23.74) [20]			
Parent Political Ideology[3]				*r*(283) = .01 *p* = .900	*r*(289) = .08 *p* = .196	*r*(172) = .16 *p* = .033
Liberal	80.27 (21.60) [183]	78.91 (23.41) [250]	77.27 (25.13) [148]			
Moderate	81.02 (22.52) [90]	83.77 (18.78) [38]	83.33 (19.25) [22]			
Conservative	86.67 (17.21) [10]	*	*			
Parent Education Level[4]				*r*(105) = .11 *p* = .258	*r*(106) = .13 *p* = .194	*r*(60) = -.03 *p* = .835
Some schooling	—	—	—			
High school diploma	*	*	*			
Some college/Associate's degree	72.55 (23.53) [17]	79.63 (20.26) [18]	75.00 (27.06) [12]			

College/Bachelor’s degree	80.44	72.69	91.67			
	(23.91) [49]	(30.38) [36]	(16.67) [20]			
Advanced degree (MA, MD, PhD, etc.)	81.20	85.95	79.63			
	(21.01) [39]	(19.54) [51]	(22.33) [27]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(285) = .08	*r*(289) = -.04	*r*(170) = -.09
Less than $25,000	80.56	93.33	82.22	*p* = .205	*p* = .510	*p* = .254
	(24.53) [6]	(11.65) [10]	(21.02) [9]			
$25,001 to $50,000	75.64	77.59	80.95			
	(24.17) [13]	(25.70) [29]	(20.52) [14]			
$50,001 to $75,000	80.00	80.46	82.35			
	(21.31) [35]	(23.18) [58]	(22.07) [34]			
$75,001 to $125,000	78.55	77.52	77.16			
	(22.94) [91]	(22.11) [86]	(23.64) [54]			
More than $125,000	82.97	80.13	76.55			
	(20.78) [138]	(23.18) [106]	(27.00) [59]			

[*] Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] Two control parents circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). These data points are excluded from the table, but were scored as “3.5” on the income scale for analyses purposes.

Outfit at appointment

Ratings of participants' outfits at appointment did not differ by race, geographic location, parent education level, parent political orientation, or household income for any of the participant groups. Table S19 shows detailed statistics for each of these analyses.

Table S19. Scores on Appointment Outfit Measure by Participant Group and Demographics

	Appointment Outfit					
	Mean (SD)			Statistics		
	Controls	Transgender	Siblings	Controls	Transgender	Siblings
Child's Race[1]				*t*(152.6) = 1.44 *p* = .151, *d* = .19	*t*(167.2) = 0.21 *p* = .837, *d* = .03	*t*(112.8) = 0.52 *p* = .607, *d* = .08
White	4.06 (0.52) [204]	4.15 (0.57) [205]	4.05 (0.58) [121]			
Non-White	4.17 (0.59) [90]	4.17 (0.55) [87]	4.09 (0.50) [53]			
Geographic Location[2]				—	*F*(5,287) = 1.32 *p* = .257, η_p^2 = .02	*F*(5,169) = 1.86 *p* = .104, η_p^2 = .05
Northeast	—	4.19 (0.51) [45]	4.10 (0.51) [24]	—		
Midwest/Upper Plains	—	4.16 (0.59) [64]	3.91 (0.64) [44]			
Southeast	—	4.04 (0.61) [48]	4.04 (0.59) [28]			
Mountain West	—	4.29 (0.45) [40]	4.25 (0.47) [29]			
Pacific Northwest	4.09 (0.55) [294]	4.23 (0.54) [54]	3.98 (0.55) [29]			
Pacific South	—	4.06 (0.62) [42]	4.23 (0.39) [21]			
Parent Political Ideology[3]				*r*(292) = - .01 *p* = .917	*r*(293) = .04 *p* = .479	*r*(175) = .03 *p* = .739
Liberal	4.09 (0.55) [186]	4.17 (0.57) [257]	4.06 (0.56) [149]			
Moderate	4.09 (0.56) [97]	4.06 (0.52) [35]	4.08 (0.55) [24]			
Conservative	4.17 (0.35) [9]	*	*			
Parent Education Level[4]				*r*(119) = -.01 *p* = .928	*r*(118) = -.04 *p* = .661	*r*(68) = -.15 *p* = .223
Some schooling	—	—	—			
High school diploma	*	*	*			
Some college/Associate's degree	4.32 (0.62) [17]	4.35 (0.52) [21]	4.23 (0.53) [14]			

College/Bachelor’s degree	4.07	4.07	4.31			
	(0.56) [57]	(0.73) [38]	(0.52) [24]			
Advanced degree (MA, MD, PhD, etc.)	4.21	4.25	4.08			
	(0.66) [45]	(0.51) [58]	(0.61) [29]			
Other	—	*	—			
More than one	—	*	*			
Household Annual Income[5]				*r*(294) = -.003	*r*(293) = -.01	*r*(175) = -.08
Less than $25,000	4.42	4.32	4.06	*p* = .959	*p* = .826	*p* = .299
	(0.58) [6]	(0.48) [11]	(0.67) [9]			
$25,001 to $50,000	4.08	4.11	4.04			
	(0.57) [13]	(0.65) [32]	(0.63) [17]			
$50,001 to $75,000	3.91	4.13	4.26			
	(0.65) [36]	(0.60) [58]	(0.43) [32]			
$75,001 to $125,000	4.18	4.21	4.03			
	(0.48) [92]	(0.52) [89]	(0.57) [51]			
More than $125,000	4.08	4.14	4.01			
	(0.55) [146]	(0.55) [103]	(0.55) [66]			

[*] Due to low sample size, unreliability of estimates, and to maintain participant confidentiality, means are not shown for any cells with fewer than five participants.

[1] Given the small number of participants within each of the minority (not White) racial groups, we collapsed those groups into a Non-White category for the statistical analyses on child’s race. Welch’s *t*-tests for unequal variances were used for comparisons.

[2] A statistic was not calculated for control group, as all participants are from the Pacific Northwest.

[3] The correlational statistics provided in this table were run using the continuous scale for parent political ideology (1 – very liberal to 7 – very conservative). The three political ideology groups presented here were created such that “Liberal” consists of those responding 1-2, “Moderate” consists of those responding 3-5, and “Conservative” consists of those responding 6-7 on the political ideology scale.

[4] The correlational statistics provided do not include those responding “Other” or “More than one” to the parent education item.

[5] One control parent circled more than one income bracket (e.g., $50,001 to $75,000 and $75,001 to $125,000). This data point is excluded from the table, but was scored as “3.5” on the income scale for analyses purposes.

Discussion

Our data yielded very few significant effects of demographic factors on gender development within these samples. More specifically, participants' gender identification, gender-typed preferences, and gender-typed behaviors did not significantly vary as a function of race (i.e., whether they were White or non-White), geographic location (i.e., what part of the USA they resided in – only applied to transgender participants and cisgender siblings), parent education level, or household income. Among cisgender controls, there were also no significant relations between their parents' political orientation and any of their gender typing measures. This was mostly true for siblings as well – the only exception being that siblings whose parents were less liberal tended to show stronger same-gender peer preferences, although this may be accounted for by a few outliers.

For transgender participants, we observed no relations between race, geographic location, parent education level and household income, and measures of gender identification or preferences. We did find a relation between parent political orientation and toy preferences. However, this was a fairly small effect (r = .17).

A major limitation in this sample is that we had very little variability in parental political orientation. The average score of all parents was very liberal (M = 1.98 on a 1 to 7 scale), with the scores in the transgender and sibling groups being especially liberal (M_{trans} = 1.66, M_{siblings} = 1.76). Therefore, to best test the possibility that children in more conservative families, and especially transgender children in more conservative families, might assert a more stereotypical gender identity or more gender-typed preferences, one would want to recruit a sample with more parents who identify as politically moderate or conservative. Only 13% of transgender children's parents identified as politically moderate parents and less than 1% as conservative. Of course, whether this effect would be stronger or would go away with the inclusion of more children from more conservative backgrounds is unclear. Further, we will note that anecdotally, some parents while completing the demographic questions spontaneously commented that their political orientation has changed as a result of having a transgender child. Therefore, parents' political orientation at this time-point may not reflect their longer-term identification as it is likely to for cisgender controls. In sum, there are many fascinating questions for future work related to the association between parents' past and current political orientation and children's gender typing, especially with families with transgender children.

Although transgender participants and cisgender controls show highly similar gender typing (see main paper), we found that there were a couple of differences between the two groups in terms of demographic characteristics. Cisgender controls, on average, had higher household income and were less liberal compared to transgender participants. One difficulty in interpreting these differences, however, is that the participants were in different geographic regions. Specifically, whereas all cisgender controls were recruited in a highly liberal city in the Pacific Northwest which was found to be the third most liberal city in the nation *(24)* and 4th-most wealthy city in the nation *(25)*, transgender participants were recruited from diverse locations throughout the U.S. and Canada varying in terms how liberal or conservative the local population is and how expensive it is to live there. Being surrounded by highly liberal or highly conservative people might influence parents' judgments of how liberal they are. For example, a parent in an extremely liberal city who identifies as a 3 on a 7-point scale of political orientation could be saying they are liberal but not as liberal as others in that city, while a parent living in a very conservative area might identify as a 2 on the same scale reasoning that they are far more liberal than the average person there. However, on key political issues (e.g., abortion), the parent with a 3 might in fact be more politically liberal than the parent who selected a 2. This is one limitation to the type of self-report measure of political orientation utilized here as well as a limitation of recruiting samples from different locations (which occurred because of feasibility constraints). In future studies it would be useful to have more "objective" questions (e.g., which candidate they voted for, how they feel about certain social issues) to assess political orientation to reduce this relative-bias issue.

The differences in income may also be explained by location biases. That is, controls may have been wealthier because they live in a more expensive than average city and therefore likely get paid more even for the same occupation. As evidence in favor of this interpretation, while the groups differed by income, they did not differ in parental education. Thus, we do not believe there is a major underlying difference in the financial security of parents of the transgender children and controls; instead, we believe that there is likely a cost-of-living difference that explains the income gap. Nonetheless, in all groups, families were wealthier on average than most American families and were therefore similar in that way.

10. Relations between parental reports of child gender development and children's self-report

Summary and take-home points. In this section, we describe registered analyses examining relations between parents' reports of their child's gender development and children's own reports. In general, the current findings suggest that parent's reports of their children's gender-typed identities and preferences are fairly well aligned with children's reports of their own identities and preferences. Further, we see yet again that the responses of transgender children (and their parents) track fairly well with the responses of cisgender controls (and their parents).

Research on gender development often focuses on children's own reports of their gender identities and their gender-typed preferences. Studies of parents' perceptions of their child's gender typing and how these perceptions might relate to children's own perceptions are rare. Whereas a few studies have examined the association between parental and child report on gender typing (*14-15*), the design used in these studies did not allow researchers to assess coherence among parents' and children's reports of children's gender typing at a single timepoint. The one area in which parent reports of gender are used more frequently is with clinical populations referred for gender-related diagnoses (e.g., Gender Identity Disorder, Gender Dysphoria). Johnson et al. *(12)* developed a parent measure of child gender identity (Gender Identity Questionnaire for Children, GIQC) that is mainly used as an assessment tool with clinical populations. The GIQC includes questions on children's gender identity and gender-typed behaviors and preferences (see *(12)* for items), and it reliably differentiates between gender-referred children and cisgender controls (*12, 26*). In a study of Canadian and Dutch children who were referred to gender clinics for Gender Identity Disorder, Wallien and colleagues (*27*) found that maternal reports on the GIQC were positively correlated with children's own reports of cross-gender identification and behavior both within children with GID (Canadian children, $r = .36, p < .001, N = 325$; Dutch children, $r = .44, p < .001, N = 210$) and within groups of control children ($r = .25, p < .001, N = 168$), and across the total sample of children ($r = .59, p < .001, N = 703$).

Together, previous research suggests that parental reports are related to children's self-reports of their identities and their gender-typed behaviors, but to date there have been few tests of this question within non-clinical samples. If parent and child reports of gender identity and gender typing correspond well with each other, this would allow researchers who want to study gender development among younger children who cannot yet express their own gender identity and preferences, to confidently rely on parent reports, as they will likely be good reflections of children's identity and preferences. Additionally, examining coherence among both identity and gender-typed preference measures could highlight potential differences in parents' awareness of their child's gender. That is, it is presumable that parents might be able to provide more insight on their child's gender-typed behaviors and preferences, than their identity, as some aspects of a child's gender identity might not be as salient as their behaviors. Thus, better understanding coherence among parent and child reports could inform future measures. The current chapter contributes to the growing but limited literature on how parents' reports might relate to children's gender identity and gender typing in cisgender and transgender children recruited through non-clinical methods.

To assess parents' perceptions of their child's gender typing, we used the GIQC developed by Johnson and colleagues (*12*). However, because our transgender children were socially transitioned, we scored it according to their lived gender (rather than according to their assigned sex). In this way, transgender girls and cisgender girls, for example, were scored identically. In addition, we asked parents what they viewed their child's current gender as, and what they thought their child's gender would be when they grew up. Finally, we also asked parents to specify whether their child's preferences for toys, clothing, and peers were more stereotypically aligned with those of girls, boys, or both/gender neutral. Due to limited prior understanding on the topic, the analyses reported in this chapter are exploratory. Thus, although no hypotheses were registered, the precise analyses that would be conducted were registered before analyses were run (https://osf.io/q2kuw/?view_only=c9f9df7d1e5a4f95ab893f28a81ce9e0).

Results

Gender Identity Questionnaire for Children (GIQC)

The descriptive statistics for parents' responses on the GIQC for each participant group are shown in Table S20. We also found that for participants who had two parents complete the GIQC, scores of the two parents were significantly correlated, r(197) = .65, p < .001. For these participants, we used an average score of both parents' responses, as registered. To examine potential differences between participant groups and genders on parents' responses to the GIQC, we conducted a 2 (child gender: boy, girl) x 3 (participant group: transgender, control, sibling) ANOVA on parents' GIQC scores. Unlike Johnson et al. *(12)*, GIQC scores were computed based on participants' genders (and not assigned sex), with higher scores indicating greater association with the child's own gender (which was also the child's sex of birth in the case of cisgender controls and siblings), and lower scores indicating greater association with the other gender (which was also the child's sex of birth in the case of transgender children). There was a significant main effect of participant group, $F(2,509) = 11.61$, $p < .001$, $\eta_p{}^2 = .04$. Post-hoc Tukey comparisons showed that parents' ratings of cisgender siblings' ($M = 4.20$, $SD = 0.34$) gender identity as more gender-typical compared to both transgender participants ($M = 4.03$, $SD = 0.43$, $p < .001$, $d = 0.44$) and cisgender controls ($M = 3.97$, $SD = 0.35$, $p < .001$, $d = 0.67$). Transgender participants were also rated by their parents as more gender-typical compared to cisgender participants, $p = .277$, $d = 0.15$. Results did not yield a significant main effect of child gender, $F(1,509) = 1.61$, $p = .206$, $\eta_p{}^2 < .01$, or a significant interaction, $F(2,509) = 2.14$, $p = .119$, $\eta_p{}^2 = .01$. To make our findings comparable to previous research for those interested, we have also calculated each participant's score based on the original scoring. With the original scoring (based on children's sex at birth), the descriptive statistics for each participant group were as follows: transgender, $M = 2.07$, $SD = 0.47$; control, $M = 3.97$, $SD = 0.35$; siblings, $M = 4.20$, $SD = 0.34$. For comparison, the scores of the transgender participants were more sex-"atypical" than the children with "gender identity disorder" in Johnson et al *(12)* whose mean score was M=2.83, SD=0.62, though the children in that study were not socially-transitioned, lived at a different time, and were being referred to a clinic specializing in the treatment of gender nonconforming children and therefore any combination of these (or other) factors may explain these differences. Interestingly, the controls (which included siblings and unrelated controls) in that paper had a mean score of 4.20 (SD=0.36), most comparable to the sibling group in our sample. The age ranges in all groups were comparable (2.5 or 3 to 12 years of age).

Table S20. Scores on Parent Measures by Child Participant Group & Age

	Age (years)									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Toy Preference[1]										
Cisgender Controls	0.62 (0.25) [*4*]	0.53 (0.47) [*24*]	0.49 (0.48) [*41*]	0.61 (0.54) [*47*]	0.55 (0.48) [*47*]	0.53 (0.53) [*35*]	0.69 (0.44) [*35*]	0.53 (0.49) [*38*]	0.80 (0.37) [*22*]	0.67 (0.56) [*23*]
Transgender	0.50 (0.63) 6*5*]	0.30 (0.54) [*20*]	0.53 (0.55) [*44*]	0.48 (0.52) [*43*]	0.53 (0.54) [*53*]	0.62 (0.52) [*34*]	0.48 (0.56) [*35*]	0.77 (0.34) [*37*]	0.68 (0.43) [*23*]	0.52 (0.55) [*20*]
Cisgender Siblings	0.78 (0.25) [*8*]	0.59 (0.35) [*14*]	0.76 (0.36) [*24*]	0.88 (0.30) [*13*]	0.80 (0.38) [*32*]	0.78 (0.36) [*30*]	0.75 (0.32) [*21*]	0.68 (0.43) [*24*]	1.00 (0.00) [*8*]	0.67 (0.43) [*15*]
Clothing Preference[1]										
Cisgender Controls	0.75 (0.29) [*4*]	0.89 (0.29) [*24*]	0.89 (0.29) [*41*]	0.73 (0.50) [*47*]	0.88 (0.32) [*47*]	0.80 (0.44) [*35*]	0.84 (0.32) [*35*]	0.84 (0.35) [*38*]	0.91 (0.25) [*22*]	0.85 (0.35) [*23*]
Transgender	1.00 (0.00) [*6*]	0.92 (0.16) [*20*]	0.91 (0.35) [*44*]	0.98 (0.09) [*43*]	0.99 (0.08) [*53*]	0.93 (0.19) [*34*]	0.97 (0.10) [*35*]	0.99 (0.08) [*37*]	0.93 (0.17) [*23*]	0.90 (0.30) [*21*]
Cisgender Siblings	0.91 (0.19) [*8*]	0.75 (0.43) [*14*]	0.98 (0.07) [*24*]	0.87 (0.30) [*13*]	0.95 (0.15) [*32*]	0.99 (0.05) [*30*]	0.95 (0.22) [*21*]	0.94 (0.22) [*24*]	1.00 (0.00) [*8*]	0.92 (0.32) [*15*]
Peer Preference[1]										
Cisgender Controls	0.00 (0.00) [*4*]	0.33 (0.56) [*23*]	0.33 (0.58) [*39*]	0.36 (0.57) [*47*]	0.50 (0.51) [*46*]	0.66 (0.48) [*35*]	0.59 (0.50) [*34*]	0.71 (0.46) [*38*]	0.64 (0.49) [*22*]	0.70 (0.56) [*23*]
Transgender	0.75 (0.42) [*6*]	0.42 (0.47) [*20*]	0.43 (0.54) [*43*]	0.56 (0.50) [*43*]	0.59 (0.47) [*52*]	0.59 (0.47) [*34*]	0.60 (0.44) [*34*]	0.64 (0.55) [*37*]	0.54 (0.58) [*23*]	0.19 (0.77) [*21*]
Cisgender Siblings	0.31 (0.37) [*8*]	0.32 (0.42) [*14*]	0.58 (0.43) [*24*]	0.62 (0.46) [*13*]	0.63 (0.48) [*31*]	0.62 (0.43) [*30*]	0.81 (0.40) [*21*]	0.69 (0.44) [*24*]	0.75 (0.46) [*8*]	0.77 (0.53) [*15*]

GIQC[2]										
Cisgender	3.75	3.87	3.95	3.89	4.02	3.91	4.01	4.00	4.12	4.06
Controls	(0.05) [*2*]	(0.40)	(0.27)	(0.41)	(0.38)	(0.36)	(0.34)	(0.30)	(0.32)	(0.35)
		[*16*]	[*24*]	[*34*]	[*31*]	[*19*]	[*25*]	[*26*]	[*14*]	[*19*]
Transgender	4.40	3.82	3.98	4.03	4.08	4.03	4.03	4.23	4.06	3.79
	(0.12) [*3*]	(0.26)	(0.58)	(0.41)	(0.43)	(0.37)	(0.38)	(0.39)	(0.35)	(0.51)
		[*12*]	[*24*]	[*29*]	[*31*]	[*16*]	[*24*]	[*25*]	[*14*]	[*16*]
Cisgender	4.25	4.12	4.26	4.21	4.15	4.24	4.33	4.10	4.50	4.04
Siblings	(0.42) [*8*]	(0.34)	(0.31)	(0.29) [*9*]	(0.31)	(0.38)	(0.26)	(0.35)	(0.21) [*6*]	(0.45)
		[*10*]	[*11*]		[*17*]	[*15*]	[*10*]	[*17*]		[*8*]
Current Gender[3]										
Cisgender	100%	96%	100%	100%	98%	100%	100%	100%	100%	100%
Controls	[*4*]	[*24*]	[*41*]	[*47*]	[*47*]	[*35*]	[*35*]	[*38*]	[*22*]	[*22*]
Transgender	50%	80%	68%	81%	81%	88%	86%	86%	83%	62%
	[*6*]	[*20*]	[*44*]	[*43*]	[*53*]	[*34*]	[*35*]	[*37*]	[*23*]	[*21*]
Cisgender	100%	100%	100%	100%	97%	100%	100%	100%	100%	100%
Siblings	[*8*]	[*14*]	[*24*]	[*13*]	[*32*]	[*30*]	[*21*]	[*24*]	[*8*]	[*15*]
Future Gender[3]										
Cisgender	100%	79%	98%	98%	100%	100%	100%	100%	95%	100%
Controls	[*4*]	[*24*]	[*41*]	[*47*]	[*47*]	[*35*]	[*35*]	[*38*]	[*22*]	[*22*]
Transgender	86%	67%	70%	84%	79%	82%	94%	95%	91%	86%
	[*6*]	[*21*]	[*44*]	[*43*]	[*53*]	[*34*]	[*35*]	[*37*]	[*23*]	[*21*]
Cisgender	100%	93%	96%	100%	97%	100%	100%	100%	100%	100%
Siblings	[*8*]	[*14*]	[*24*]	[*13*]	[*32*]	[*30*]	[*21*]	[*24*]	[*8*]	[*15*]

Note. Means, standard deviations (in parentheses), and *n*s (in brackets) for parent report of child's Toy Preference, Clothing Preference, and Peer Preference as well as scores on the GIQC presented for each participant group at each age. Percentage of parents providing their child's gender in response to the Current and Future Gender Identity items and total *n*s (in brackets) are presented for each group at each age. There were two 13-year-old cisgender control participants whose data was included with the 12-year-old data.

[1] Responses were scored with '1' point if parents selected the child's gender, '-1' points if they selected the other gender, '0' points if they selected the neutral response or if they selected both genders.

[2] Responses to the GIQC were coded such that a score of '5' would indicate responses most aligned with the child's gender, and a score of '1' would indicate responses most aligned with the other gender (for transgender participants, a score of '1' indicated responses most aligned with the child's assigned sex).

[3] Percentages represent the percent of parents providing their child's identified gender.

To examine the extent to which parents' responses on the GIQC corresponded to participants' gender identity and gender-typed preferences, we conducted separate correlation analyses for each participant group between the parent GIQC scores and children's scores on each child measure. Results for each group are shown in Table S21-S23 below. As indicated by the significant positive correlations, overall, parents' evaluation of their child's gender identity and preferences were consistent with children's reports. However, there were a few exceptions within each participant group and for particular measures. Participants' scores on the IAT did not significantly correlate with the GIQC for controls and siblings: control, *r*(149) = -.10, *p* = .241; siblings, *r*(64) = .07, *p* = .589. However, there was a significant correlation between scores on the IAT and GIQC for transgender participants, *r*(124) = .19, *p* = .039. It is important to note, however, that sample sizes for these analyses, particularly those for the sibling group, was low, making these results less reliable. Additionally, for control participants, parents' GIQC scores did not show a significant correlation with children's explicit gender identity, *r*(208) = .09, *p* = .194. Similarly, for siblings, parents' GIQC scores did not show significant correlations with children's explicit gender identity, *r*(110) = .04, *p* = .694, their perceptions of similarity to own vs. other gender, *r*(96) = .18, *p* = .076, and their peer preferences, *r*(102) = .07, *p* = .490, though again these involved smaller sample sizes than most of the other analyses. As a reminder, we did generally find a restricted range for the explicit gender identity items so this may have contributed to the failure to find effects for that measure. For transgender participants, all measures other than the IAT were significantly and positively correlated with the GIQC.

Table S21. Correlation Matrix of Child Measures with GIQC Scores for Cisgender Control Participants

Cisgender Control Participants		
	GIQC Score	***p*-value**
1. Toy Preference	***r*(190) = .38**	***p*<.001**
2. Clothing Preference	***r*(190) = .25**	***p*=.001**
3. Peer Preference	***r*(183) = .30**	***p*<.001**
4. Similarity[1]	***r*(189) = .39**	***p*<.001**
5. Gender Identity IAT	*r*(149) = -.10	*p*=.241
6. Appointment Outfit	***r*(210) = .29**	***p*<.001**
7. Explicit Gender Identity[2]	*r*(208) = .09	*p*=.194

Note. Variables are coded such that higher values represent responses more in line with participants' own gender.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender (own gender similarity minus other gender similarity).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with other gender twice) to 2 (responded with own gender twice).

Table S22. Correlation Matrix of Child Measures with GIQC Scores for Transgender Participants

Transgender Participants		
	GIQC Score	***p*-value**
1. Toy Preference	***r*(178) = .42**	***p*<.001**
2. Clothing Preference	***r*(178) = .22**	***p*=.003**
3. Peer Preference	***r*(174) = .15**	***p*=.049**
4. Similarity[1]	***r*(172) = .40**	***p*<.001**
5. Gender Identity IAT	***r*(124) = .19**	***p*=.039**
6. Appointment Outfit	***r*(193) = .18**	***p*=.012**
7. Explicit Gender Identity[2]	***r*(192) = .24**	***P*<.001**

Note. Variables are coded such that higher values represent responses more in line with participants' own gender.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender (own gender similarity minus other gender similarity).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with other gender twice) to 2 (responded with own gender twice).

Table S23. Correlation Matrix of Child Measures with GIQC Scores for Cisgender Sibling Participants

Cisgender Sibling Participants		
	GIQC Score	***p*-value**
1. Toy Preference	*r*(103) = .22	*p*=.027
2. Clothing Preference	*r*(103) = .22	*p*=.027
3. Peer Preference	*r*(102) = .07	*p*=.490
4. Similarity[1]	*r*(96) = .18	*p*=.076
5. Gender Identity IAT	*r*(64) = .07	*p*=.589
6. Appointment Outfit	*r*(111) = .20	*p*=.038
7. Explicit Gender Identity[2]	*r*(110) = .04	*p*=.694

Note. Variables are coded such that higher values represent responses more in line with participants' own gender.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender (own gender similarity minus other gender similarity).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with other gender twice) to 2 (responded with own gender twice)

Parent and child report on child gender identity

For each participant group, we report the percentage of parents who responded with their child's current gender (as determined by their current pronoun use) for their current and future identity in Table S20. To examine whether parents' reports on child gender identity was consistent with children's own explicit reports of their gender identity, we conducted separate correlation analyses between the parent composite score (i.e., the sum of their responses to their child's current and future identity) and the child composite score (i.e., the sum of their responses to their current and future identity) for each participant group. We found that although there was a significant, though small, correlation between parent and child report of gender identity for transgender children, *r*(308) = .12, *p* = .040, there were no significant correlations for cisgender controls, *r*(312) = -.05, *p* = .381, and siblings, *r*(188) = .014, *p* = .846. In general, across all three groups the magnitude of these effects were quite small. To understand these findings further, we cross-tabulated parents' and children's responses, and found that the variation in scores was very low, with an especially low range of responses for parents reporting on controls and siblings. For example, whereas 97% of control parents and 98% of parents of siblings provided a score of '2' (i.e., responding with child's own gender for both current and future gender), 71% of parents of transgender children reported '2'. Examined separately for current and future identity questions, we found that 99% of parents of cisgender controls and 99% of cisgender siblings responded with their child's gender for their current gender (in contrast to 79% of parents of transgender children), and 97% of parents of controls and 98% of parents of siblings responded with their child's gender for their future gender (in contrast to 83% of parents of transgender children). These scores meant that there was functionally no variability in parent scores to predict child scores for controls and siblings. Even for transgender children there was remarkably little variability. Thus, the lack of a significant relationship in 2 cases, and the very small correlation in the third, is due to restriction of range. If instead of using a correlation (as we registered), we examine the percentage of parents and children with absolute agreement, we see very high rates of agreement (i.e., 71% of controls, 80% of siblings and 60% of transgender parents and children gave the exact same responses on these items).

Parent and child report on child gender-typed preferences

We conducted separate correlations for each participant group, between parents' reports of their children's toy, clothing and peer preferences, and children's own reports of their toy, clothing and peer preferences. For transgender and control participants, all three measures were significantly and positively correlated. For siblings, although parent and child reports of clothing preferences were significantly and positively correlated, there were no significant correlations for toy and peer preferences. Results are shown in Table S24 below.

Table S24. Correlation Matrix of Child Measures with Parent Measures within each group of participants

	Measure	**Statistic**	***p*-value**
Cisgender Control Participants	**Toy Preference**	**_r_(274) = .33**	**_p_<.001**
	Clothing Preference	**_r_(274) = .34**	**_p_<.001**
	Peer Preference	**_r_(281) = .19**	**_p_=.001**
Transgender Participants	**Toy Preference**	**_r_(274) = .47**	**_p_<.001**
	Clothing Preference	**_r_(274) = .27**	**_p_<.001**
	Peer Preference	**_r_(286) = .19**	**_p_=.002**
Cisgender Sibling Participants	**Toy Preference**	*r*(163) = .14	*p*=.072
	Clothing Preference	**_r_(163) = .33**	**_p_<.001**
	Peer Preference	*r*(172) = .12	*p*=.105

Note. Variables are coded such that higher values represent responses more in line with participants' own gender.

Discussion

One major take-home from these findings is that there is a clear and fairly consistent relation between parent reports of gender-typed behavior and children's actual gender-typed behavior even in non-clinically recruited samples—the first test we know of at a single time point in a non-clinical sample. When we found significant variability in responding we tended to find an association suggesting that parents are fairly good reporters of their children's gender-typed behaviors. Importantly, parents were not present during children's responses, meaning they could not have used responses during the session itself to inform their reports.

The coherence among parents' and children's reports of gender identities and preferences was apparent across several of the measures used and was particularly pronounced for transgender participants and cisgender controls. Overall, parents' responses on the GIQC were positively correlated with children's identity and preference measures across all three groups. Similarly, parents' reports of their child's gender-typed toy, clothing and peer preferences were all positively correlated with their children's reports in the case of transgender participants and cisgender controls. Parents' reports of their children's gender identity were only correlated with children's reports of gender identity in the case of transgender children. However, we believe the lack of association for cisgender controls (and siblings) on this measure was driven by a lack of variance in parents' responses (they nearly always assumed these children were and would remain cisgender).

One unexpected pattern was that we found parents to be less good reporters of cisgender siblings' gender typing. On the GIQC, parents reported more gender-typed identities and preferences for cisgender siblings compared to transgender participants and controls, and they reported more gender-typed identities and preferences for transgender participants than controls. This is inconsistent with children's own reports of their gender typing – specifically, siblings did not report or appear more gender typed than their transgender siblings or cisgender controls. Additionally, though overall, parents' responses on the GIQC were positively correlated with children's identity and preference measures for all three groups, the magnitude of these effects were less pronounced for siblings. As with the GIQC, parents' reports on their child's gender-typed preferences only correlated with their cisgender children's (i.e., the sibling group) when asked about clothing preferences, and not for toy or peer preferences. These findings are surprising because these are the exact same parents who were more accurate in reporting their transgender children's responses. Parents' differential responding for their transgender child and cisgender sibling might be due to the fact that, during appointments, parents often complete the transgender questionnaires followed by the sibling questionnaires (this was done procedurally in case time ran out and parents didn't have time to do both forms; the forms of transgender participants were deemed more valuable due to difficulties in recruitment of transgender participants). Perhaps completing the form later in the session when they were less on-task or more tired, led to more inaccuracies. Other possibilities are that within the session (again, because of the order), or just in general, parents may have (intentionally or unintentionally) completed the sibling questionnaires thinking of this child relative to their transgender child or parents of transgender children may think about their transgender child's gender so much that it crowds out their attention to the sibling's gender behaviors. For either of these reasons parents may simply think of the sibling as "pretty average for a girl (or boy)" and respond according to gender stereotypes.

Additionally, our results show that parents of transgender children and their siblings responded differently on the GIQC compared to parents of controls, with the former group of parents viewing their children's identity and behaviors as more gender-typical than the latter. One reason for this might be that parents of transgender children and cisgender siblings are more likely to perceive their children's behaviors as more gender-normative than they actually are.

Alternatively, or at the same time, parents of cisgender controls might be more likely to perceive their children's behaviors as more gender-atypical than they actually are.

An additional overall finding was that, in their responses on the GIQC, parents rated boys as more gender-typed than girls, which is consistent with our findings described in the main text, where boys tended to show more gender-typed identities and preferences compared to girls. We suspect this finding can be explained by North American norms wherein a wider range of preferences and behaviors are tolerated for girls than boys (for a review see, *28)*.

11. Registered analyses repeated for participants not previously reported in publications before registration

Data from a total of 177 (transgender participants, N = 73) out of our 822 total participants have been previously reported in one or more of four publications that was published before we began working on this manuscript (1-4; see Table S2 for exact numbers of participants and tasks featured in each previous study, and for exact numbers of participants included in current analyses by group). All analyses in this section follow the registered analysis plan for child measures (https://osf.io/q2kuw/), excluding participants featured in previously published papers. Participants were only excluded from those analyses for which their data had been previously reported. For example, if only a participant's explicit gender identity had been reported in a previous publication, then they were only excluded from analyses involving explicit gender identity, and not excluded from other analyses of other measures. We report the results of each of our tasks measuring gender identity and gender typed preferences, including measures of explicit gender identity, implicit gender identity, similarity to own- and other-gender children, toy preferences, clothing preferences, peer preferences, and outfit at appointment. For each of these tasks, we make comparisons by participant group and participant gender and examine the relationships between our tasks and time since transition and age. Overall, the results described here are highly consistent with those from the whole sample. The exclusion of these participants had no impact on the significance of any of the findings—all significant results remained significant, while all non-significant results remained non-significant—indicating that the participants included in previous research did not influence the findings.

Measures of gender identity

Explicit gender identity

Comparisons by participant group. Participants could indicate whether their current and future gender identities aligned with their current lived gender (i.e., a girl said she was a girl), aligned with the other gender (i.e., a girl said she was a boy), or an alternative response (e.g., a girl said she was both or neither a boy and a girl). Due to the very low number of participants who selected the "other gender" option (current gender identity: n = 2 transgender, n = 1 siblings, n = 3 controls; future gender identity: n = 4 transgender, n = 2 siblings, n = 2 controls), we conducted Fisher's exact tests.

We conducted two separate Fisher's exact tests, one for current gender identity, and one for future gender identity, comparing transgender, control and sibling participants' likelihood to provide own-gender, other-gender, or alternative responses. As shown in the main text, we found that participant groups did not differ in their responses about their current gender identity (p = .680, V = .04) or their future gender identity (p = .258, V = .06). Overall, children were still highly likely to respond to the current and future identity questions by asserting their current gender identity (86% and 83% respectively).

Comparisons by participant gender. Two Fisher's exact tests were used for the same reasons described above, to examine potential gender differences in participants' responses to the current and future gender identity questions. Consistent with results from the main text, results showed that girls and boys differed in their responding when asked about their current gender identity (p < .001, V = .16) and when asked about their future gender identity (p < .001, V = .14). Both groups still showed a tendency to respond with their current gender for both the current (81% of girls; 93% of boys) and future (79% of girls, 89% of boys) identity questions, with the difference between genders being driven by 18% of girls and 7% of boys on the current identity question and 20% of girls and 9% of boys on the future identity question replying with a response that was neither boy nor girl.

Comparisons to chance. We conducted chi-square goodness of fit tests, separate for the current gender identity and future identity questions, to assess whether participants' responses differed from chance. We conducted separate analyses for girls and boys because there were gender differences in our previous analyses. Again, both boys' (*p*s < .001) and girls' (*p*s < .001) responses differed from chance responding, meaning they were more likely to give some answers than others. Examination of the response percentages indicated that the overwhelming number of participants gave their current gender as their current (93% of boys, 81% of girls) and future (89% of boys, 79% of girls) identity.

Relation between current and future identity, and age-related changes. As in the main text, these two identity items were still strongly correlated, $r(661) = .51, p < .001$.

To examine whether participants' responses to the explicit gender identity questions changed with age, we added together the scores of the current and future identity questions and correlated this composite with age. Consistent with results from the main text, we found a significant positive correlation, $r(661) = .18, p < .001$, indicating that explicit identification with own gender increased with age.

Implicit gender identity

Group comparisons. We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' Gender Identity IAT scores. Consistent with our findings from the main paper, we found a significant main effect of participant group, $F(2,457) = 4.95, p = .008, \eta_p^2 = .02$. Follow-up post-hoc Tukey HSD comparisons still showed that transgender participants ($M = .24, SD = .45$) scored lower on the IAT compared to control participants ($M = .37, SD = .48, p = .018, d = .28$) and siblings ($M = .40, SD = .44, p = .013, d = .36$), indicating that controls and siblings showed stronger implicit identification with their own gender than transgender participants (though in the main paper, siblings did not differ significantly from transgender participants). Also consistent with results from the whole sample, there were no gender differences, $F(1,457) = 2.44, p = .119, \eta_p^2 = .01$, and no interaction, $F(2,457) = 1.15, p = .316, \eta_p^2 = .01$.

Comparisons to chance. We conducted separate one-sample *t*-tests for each participant group because there was a significant group difference, comparing IAT scores to chance (0). Transgender participants, $t(171)= 7.04, p < .001, d = 0.54$, controls, $t(191) = 10.76, p < .001, d = 0.78$, and siblings, $t(98) = 9.07, p < .001, d = 0.91$, were all still significantly above the neutral response on the IAT, indicating that participants in all groups identified more with their own gender than the other gender. Participants' implicit own-gender identification for the overall sample was also still significantly above chance, $t(464) = 15.28, p < .001, d = 0.71$.

Age-related change. We conducted correlation analyses between age and implicit identification. We again did not find a significant correlation, $r(463) = .04, p = .372$.

Similarity to own- and other-gender children

Similarity to own gender. We conducted separate 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVAs on participants' perceived similarity to their own gender and the other gender. Consistent with our findings from the main paper, there was not a significant main effect of participant group on similarity to one's own gender, $F(2,623) = 1.97, p = .141, \eta_p^2 = .01$. We again found a significant main effect of participant gender on perceived similarity to own gender, $F(1,623) = 31.56, p < .001, \eta_p^2 = .05$, indicating that boys ($M = 4.36, SD = .67$) felt greater similarity to their own gender than girls did ($M = 4.02, SD = .88$). Finally, we still did not find a significant participant group x gender interaction, $F(2,623) = 2.17, p = .115, \eta_p^2 = .01$.

Similarity to other gender. Our findings for similarity to other gender showed the same pattern of results as displayed in the main text. There was not a significant main effect of participant group: $F(2,622) = 0.44, p = .642, \eta_p^2 < .01$. We still found a significant main effect of participant gender, $F(1,622) = 11.63, p < .001, \eta_p^2 = .02$, suggesting that boys ($M = 1.92, SD = .77$) perceived lower similarity to the other gender than girls did ($M = 2.19, SD = .87$). Finally, we found no significant participant group x gender interaction: $F(2,622) = 0.36, p = .699, \eta_p^2 < .01$.

Comparisons to the midpoint of scale. We calculated a difference score for each participant, subtracting perceived similarity to the other gender from perceived similarity to own gender to indicate whether children felt more similar to their own gender than the other gender. One-sample *t*-test comparisons of this difference score to the midpoint of the scale (0), separate for each gender, were conducted given the gender differences found in the ANOVAs reported above. As in the main text, we found that both girls, $t(378) = 26.85$, $p < .001$, $d = 1.38$, and boys, $t(248) = 34.19$, $p < .001$, $d = 2.17$, show greater perceived similarity to their own gender than the other gender. We also conducted one-sample *t*-tests for the overall sample, and found again that participants rated themselves as more similar to their own gender than the other gender, $t(627) = 40.40$, $p < .001$, $d = 1.61$.

Age-related change. We again conducted correlation analyses between age and the difference score. There was still not a significant correlation, $r(628) = .06$, $p = .112$, indicating no age-related changes.

Relations between similarity to own and other gender. We also conducted correlations between perceived similarity to own and other gender. We still found a negative correlation, $r(628) = -.19$, $p < .001$, suggesting that in general the more a child felt similar to their own gender the less similar they felt to the other gender.

Measures of gender-typed preferences

Toy preferences

Group comparisons. We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' toy preferences. Similar to results from the whole sample, there was not a significant main effect of participant group, $F(2,630) = 0.21$, $p = .813$, $\eta_p^2 < .01$. We again found a significant main effect of participant gender, $F(1,630) = 71.80$, $p < .001$, $\eta_p^2 = .10$, indicating that boys' toy preferences ($M = 77.41$, $SD = 14.98$) were more stereotypically masculine than girls' preferences were stereotypically feminine ($M = 63.07$, $SD = 22.15$). Finally, there again was not a significant interaction between participant group x participant gender, $F(2,630) = 0.35$, $p = .708$, $\eta_p^2 < .01$.

Comparisons to gender-neutral preferences. Again, given the significant gender differences and lack of a significant effect of participant group, we conducted a series of one-sample *t*-test comparisons to gender-neutral preferences (midpoint of scale = 50) for each gender group (collapsed across participant group). As in the main text, we found that both boys, $t(250) = 28.98$, $p < .001$, $d = 1.83$, and girls, $t(384) = 11.58$, $p < .001$, $d = 0.59$, showed preferences for toys stereotypically associated with their own gender. These findings were also again consistent with the overall sample, $t(635) = 22.67$, $p < .001$, $d = 0.90$.

Age-related change. We conducted correlation analyses to examine the relation between age and gender-typed toy preferences. We again found no significant relationship, $r(636) = -.007$, $p = .866$, indicating that children's preferences for gender-typed toys did not change with age.

Clothing preferences

Group comparisons. We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' clothing preferences. As found in the whole sample, there was a significant main effect of participant group, $F(2,630) = 6.93$, $p = .001$, $\eta_p^2 = .02$. Post-hoc Tukey comparisons showed that transgender participants' ($M = 86.98$, $SD = 16.64$) preferences in clothing were more gender-typed compared to controls ($M = 81.85$, $SD = 17.54$, $p = .003$, $d = .30$) and siblings ($M = 82.38$, $SD = 17.48$, $p = .027$, $d = .27$); the latter two groups did not differ from each other ($p = .952$). Different from the main analyses, we did not find a significant main effect of participant gender, $F(1,630) = 3.70$, $p = .055$, $\eta_p^2 = .01$, indicating that boys' clothing preferences ($M = 82.03$, $SD = 15.45$) were no less stereotypically masculine than girls' preferences were stereotypically feminine ($M = 85.19$, $SD = 18.36$). There was still not a significant interaction of participant group x participant gender, $F(2,630) = 1.37$, $p = .255$, $\eta_p^2 < .01$.

Comparisons to gender-neutral preferences. We conducted a series of one-sample *t*-test comparisons to gender neutral clothing preferences (midpoint of scale = 50) for each subgroup (transgender girls, transgender boys, control girls, control boys, sibling girls, sibling boys) as there were significant gender and participant group differences. In line with the results of the main paper, we found that all groups were above the midpoint in their preferences for clothing stereotypically associated with their own gender: transgender girls, *t*(160) = 25.81, *p* < .001, *d* = 2.03; transgender boys, *t*(82) = 25.73, *p* < .001, *d* = 2.82; control girls, *t*(158) = 23.36, *p* < .001, *d* = 1.85; control boys, *t*(83) = 16.65, *p* < .001, *d* = 1.82; sibling girls, *t*(64) = 14.54, *p* < .001, *d* = 1.80; sibling boys, *t*(83) = 17.32, *p* < .001, *d* = 1.89. Across the three groups, participants were again significantly more likely to prefer own-gender-typed clothing, compared to gender-neutral clothing, *t*(635) = 49.40, *p* < .001, *d* = 1.96.

Age-related change. We again conducted correlations between participant age and clothing preferences, and still found a significant negative correlation, *r*(636) = -.17, *p* < .001, indicating that participants' preferences in gender-typed clothing decreased with age.

Peer preferences

Group comparisons. We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on participants' peer preferences. As with the whole sample, there was not a significant main effect of participant group, *F*(2,591) = 0.45, *p* = .637, η_p^2 < .01. There was still a significant main effect of participant gender, *F*(1,591) = 8.43, *p* = .004, η_p^2 = .01, indicating that boys (*M* = 76.80, *SD* = 24.61) showed lower same-gender peer preferences compared to girls (*M* = 82.11, *SD* = 21.00). Additionally, there was still not a significant interaction of participant group x participant gender, *F*(2,591) = 0.92, *p* = .401, η_p^2 < .01.

Comparisons to chance. We conducted a series of one-sample *t*-test comparisons to chance (chance = 50) for each gender group (collapsed across participant group) due to the significant gender differences and lack of a significant effect of participant group. Consistent with the findings from the main paper, we found that both boys, *t*(229) = 16.52, *p* < .001, *d* = 1.09, and girls, *t*(366) = 29.29, *p* < .001, *d* = 1.53, preferred same-gender peers. We, again, found these same patterns of preferring same-gender peers in the overall sample, *t*(596) = 32.52, *p* < .001, *d* = 1.33.

Age-related change. We still did not find a significant correlation between participant age and peer preferences, *r*(597) = .004, *p* = .928.

Measures of gender-typed behavior

Outfit at appointment

Group comparisons. We conducted a 3 (participant group: transgender, controls, siblings) x 2 (participant gender: boy, girl) ANOVA on ratings of participants' outfits at appointments. There was still no significant main effect of participant group, *F*(2,651) = 1.68, *p* =.188, η_p^2 = .01, and a significant main effect of participant gender, *F*(1,651) = 40.60, *p* < .001, η_p^2 = .06, indicating that boys' (*M* = 4.27, *SD* = .47) outfits were more stereotypically masculine than girls' outfits were stereotypically feminine (*M* = 3.99, *SD* = .58). This main effect was again qualified by a significant participant group x gender interaction, *F*(2,651) = 3.31, *p* = .037, η_p^2 = .01. Simple effects analyses showed that, among all participants, boys' outfits at appointment (*M*transgender = 4.28, *M*control = 4.36, *M*sibling = 4.16) were rated as more stereotypically masculine than girls' outfits (*M*transgender = 4.06, *M*control = 3.92, *M*sibling = 3.98) were rated as stereotypically feminine (transgender: *p* = .003; controls: *p* < .001; siblings: *p* = .029).

Comparisons to gender neutral outfits. We conducted one-sample *t*-test comparisons to gender neutral clothing (midpoint of scale = 3) separately for boys and girls in each participant group due to the significant interaction effect. As in the main text, girls and boys in all groups were more likely to wear outfits associated with their own gender, compared to gender neutral (transgender boys: *t*(85) = 26.60, *p* < .001, *d* = 2.87; transgender girls: *t*(161) = 22.30, *p* < .001, *d* = 1.75; control boys: *t*(87) = 28.37, *p* < .001, *d* = 3.02; control girls: *t*(162) = 21.52, *p* < .001, *d* =

1.69; sibling boys: $t(88) = 22.45$, $p < .001$, $d = 2.38$; sibling girls: $t(68) = 14.42$, $p < .001$, $d = 1.74$). We again found the same pattern of wearing own-gender-typed clothing more than gender neutral clothing in the overall sample, $t(656) = 51.07$, $p < .001$, $d = 1.99$.

Age-related change. We conducted correlations between age and outfit at appointment as we did in chapter 3, and again found a significant negative correlation, $r(657) = -.19$, $p < .001$, indicating that participants' outfit ratings became less gender-typed as they grew older.

Time since social transition

We were again interested in whether the amount of time since transgender participants socially-transitioned impacted their gender typing. Due to the correlation between time since transition and age, $r(317) = .29$, $p < .001$, we conducted a series of partial correlations between time since their social transition and their scores on each of the measures reported above, while controlling for participant age. Results still showed no significant relations between time since transition and any of the measures after controlling for age (see Table S25, below, for statistics).

Table S25. Child Measures Correlated with Time Since Transition for Transgender Participants

	Time Since Transition
Toy Preference	*r*(244) = -.10, *p* = .251
Clothing Preference	*r*(244) = -.08, *p* = .219
Peer Preference	*r*(229) = -.04, *p* = .568
Similarity[1]	*r*(240) = -.04, *p* = .564
Gender Identity IAT	*r*(172) = .10, *p* = .236
Appointment Outfit	*r*(248) = -.02, *p* = .787
Explicit Identity Composite[2]	*r*(250) = .04, *p* = .553

Note. Partial correlations, controlling for participant age, among participants who were not reported in previous papers.

[1] The similarity variable used here is the difference between similarity to own gender and similarity to other gender (own gender similarity minus other gender similarity).

[2] The explicit identity variable used here is a composite of the current gender identity and future gender identity items. Scores range from -2 (responded with other gender twice) to 2 (responded with own gender twice

12. Equivalence tests for participant group differences

After running all analyses, our null hypothesis testing approach had indicated a lack of significant difference between transgender participants and both comparison groups. However, traditional null hypothesis testing does not allow us to test the null hypothesis. For this reason, we completed post-hoc equivalence tests, using the TOST procedure (*29)*. We became aware of these analyses after the full registered analyses had been completed. All comparisons in Table S26, ask whether the difference between the groups is less than a small effect (d = .2). This value was selected because in our registration, we specified an interest in effects larger than d = .2. We ran these analyses only for measures where there were no significant group differences in our main analyses.

Table S26. Equivalence tests for participant group differences

	Transgender vs. Control	Transgender vs. Siblings
Toy preferences	**t(543.62)=1.70, p = .045**	t(366.13)=0.26, p = .396
Peer preferences	**t(573.12)=1.69, p = .045**	t(343.26)=1.53, p = .064
Similarity to own gender	t(517.66)=1.01, p = .156	t(313.47)= 1.3, p = .098
Similarity to other gender	**t(523.35)= 1.90, p = .029**	t(328.48)=1.08 , p = .140
Explicit gender identity	**t(614.72)=2.49, p=.006**	t(423.74)=0.87, p=.193

Note. To confirm that the null findings from participant group comparisons in the main paper reflected likely equivalence, we conducted the two one-sided t-test procedure (TOST; *(29)*), which yields Welch's t-test values to show whether a difference between two groups reflects a small, or near-zero, effect (*d*).

13. Participant photos as a proxy for early socialization

Throughout the paper and in past work, we have operated under the assumption that transgender children were initially reared according to the gender traditionally associated with their sex at birth. At various points people have questioned this assumption in our work and in general. Therefore, we began asking families with transgender participants from our study to provide photographs of their children's first three years. The idea was that photographs could provide an indirect yet relatively objective measure of early socialization. Prior to testing, parents were given a list of 15 photos (child's first three birthday parties, nursery upon birth, infancy crib, bedroom at first three ages, the first day the child was brought home from the hospital, child's first three Halloween costumes, family holiday photos from the child's first three years) and were asked to share as many as possible, if they and their children felt comfortable with doing so (as with any other aspect of their participation, this was completely voluntary). We provided this list to ensure photos would depict similar events across participants and because we believed at such occasions parents often select what children will wear. During the testing session, the experimenter occluded any faces or identifying information present in the photos, labeled the photos with the participant number for ease of matching to participant group, and took photos of the photos. We received photos for most families we visited during the time in which we collected photographs. The final set included 98 transgender children, and on average, families provided 9 photos each (often they reported not having photos from a given event because, for example, they had gotten lost or another child was born and therefore they did not take children trick or treating, they moved and lost an album, etc.).

Photos were coded by two independent raters who did not know that they were coding photos of only transgender participants and they did not know the child's sex assigned at birth. The coding scheme was identical to that used in coding participants' outfit at appointment (1: most stereotypically masculine, 5: most stereotypically feminine). Prior to coding, raters were trained on a separate set of photos. Once each rater completed their coding, we calculated inter-rater reliability on the whole set (930 photos), and found high reliability, $\alpha = .91$. For analysis purposes, and because families differed in the number of photos they provided, we calculated a mean score for each participant, and used these scores in one-sample t-test comparisons to the mid-point of the scale (3), which indicates gender-neutral appearance. We found that transgender girls (i.e., assigned males) were significantly more masculine in personal and room appearance ($M = 2.53$, $t(68) = -8.17$, $p < .001$), and transgender boys (i.e., assigned females) were significantly more feminine in personal and room appearance ($M = 3.54$, $t(28) = 5.98$, $p < .001$), when compared to gender-neutral; in addition the difference between transgender girls and boys was significant, $F(1,96) = 90.95$, $p < .001$. These findings are consistent with our initial assumption that the transgender children in our study were initially reared in line with their assigned sex.

Acknowledgements
We would like to thank all participating families, as well as those who have helped with data collection and data entry: Madeleine DeMeules (who also assisted with materials and subject recruitment), Samantha Barney, Gabrielle Lindquist (who also assisted with materials and subject recruitment), Deja Edwards, Melanie Fox, Arianne Eason, Allison Skinner, Vincenza Standridge, Haley Doyle, Ali Liao, Zeena Rivera, Joey Lu, Emily Gill, Mackenzie Cummings, Genna Kieper, Lindsey Wolfram, Marisa McCannn, Tayla Bolden, Sarah Gage, Alice Xeuting Li, Jenna Weber, Hanna Peterson, Grace Kim, Elaine Currie, Martina Khurana, Megan Baffaro, Bea Taylor, Michelangelo Pegasus, Sarah Colombo, Sara Duffy, Dayna Bennett, Natalie Yocum, Ashley Nowain; This research was funded by grants from the National Science Foundation (BCS-1715068 to Gülgöz and BCS-1523632 and SMA-1837857 to Olson), the National Institute of Child Health and Human Development (HD092347 to Olson), and the Arcus Foundation.

References for SI Appendix citations

1. K. R. Olson, A. C. Key, N. R. Eaton, Gender cognition in transgender children. *Psychol. Sci*. **26**, 467-474 (2015).
2. K. R. Olson, E. A. Enright, Do transgender children (gender) stereotype less than their peers and siblings? *Dev. Sci*. **21**, e12606 (2018).
3. A. A. Fast, K. R. Olson, Gender development in transgender preschool children. *Child Dev*. **89**, 620-637 (2018).
4. J. R. Rae, K. R. Olson, Test-retest reliability and predictive validity of the Implicit Association Test in children. *Dev. Psychol*. **54**, 308-330 (2017).
5. D. B. Hill, E. Menvielle, K. M. Sica, An affirmative intervention for families with gender variant children: Parental ratings of child mental health and gender. *J. Sex Marital Ther*. **36**, 6-23 (2010).
6. P. T. Cohen-Kettenis, A. Owen, V. G. Kaijser, S. J. Bradley, K. J. Zucker, Demographic characteristics, social competence, and behavior problems in children with gender identity disorder: A cross-national, cross-clinic comparative analysis. *J Abnorm Psychol*. **31**, 41-53 (2003).
7. J. Henrich, S. Heine, A. Norenzayan, The weirdest people in the world? *Behav. Brain Sci*. **33**, 61-83 (2010).
8. D. Shumer, J. Carswell, oral abstract presented at the WPATH 24th Scientific Symposium, Amsterdam, 19 June, 2016.
9. C. Smith, "Building a family: Is going into debt for in vitro or adoption worth it?" (HuffPost, 2013; https://www.huffingtonpost.com/carrie-smith/building-a-family-is-goin_b_3304149.html). [the easiest access to this source is via the URL]
10. C. L. Martin, N. C. Andrews, D. E. England, K. Zosuls, D. N. Ruble, A dual identity approach for conceptualizing and measuring children's gender identity. *Child Dev*. **88**, 167-182 (2017).
11. A. G. Greenwald, B. A. Nosek, M. R. Banaji, Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol*. **85**, 197-216 (2003).
12. L. L. Johnson, S. J. Bradley, A. S. Birkenfeld-Adams, M. A. R. Kuksis, D. M. Maing, J. N. Mitchell, K. Zucker, A parent-report gender identity questionnaire for children. *Arch. Sex. Behav*. **33**, 105-116 (2004).
13. D. N. Ruble, L. J. Taylor, L. Cyphers, F. K. Greulich, L. E. Lurye, P. E. Shrout, The role of gender constancy in early gender development. *Child Dev*. **4**, 1121-1136 (2007).
14. J. E. O. Blakemore, 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 (2003).
15. M. L. Signorella, R. S. Bigler, L. S. Liben, Developmental differences in children's gender schemata about others: A meta-analytic review. *Dev. Rev.* **13**, 147-183 (1993).
16. H. M. Trautner, D. N. Ruble, L. Cyphers, B. Kirsten, R. Behrendt, P. Hartmann, Rigidity and flexibility of gender stereotypes in childhood: Developmental or differential? *Infant Child Dev.* **14**, 365-381 (2005).
17. M. L. Halim, D. N. Ruble, C. S. Tamis-LeMonda, K. M. Zosuls, L. E. Lurye, F. K. Greulich, Pink frilly dresses and the avoidance of all things "girly": Children's appearance rigidity and cognitive theories of gender development. *Dev. Psychol*. **50**, 1091-1101 (2014).
18. L. A. Serbin, K. K. Powlishta, J. Gulko, C. L. Martin, M. E. Lockheed, The development of sex typing in middle childhood. *Monogr. Soc. Res. Child Dev*. **58**, 1-74 (1993).
19. J. M. Bailey, K. T. Bechtold, S. A. Berenbaum, Who are tomboys and why should we study them? *Arch. Sex. Behav*. **31**, 333-341 (2002).

20. D. Cvencek, M. Kapur, A. N. Meltzoff, Math achievement, stereotypes, and math self-concepts among elementary-school students in Singapore. *Learn. Instr.* **39**, 1-10 (2015).
21. M. Weinraub, L. P. Clemens, A. Sockloff, T. Ethridge, E. Gracely, B. Myers, The development of sex role stereotypes in the third year: Relationships to gender labeling, gender identity, sex-typed toy preference, and family characteristics. *Child Dev*. **55**, 1493-1503 (1984).
22. L. Kulik, The impact of social background on gender-role ideology: Parents' versus children's attitudes. *J. Fam. Issues*. **23**, 53-73 (2002).
23. G. D. Levy, Relations among aspects of children's social environments, gender schematization, gender role knowledge, and flexibility. *Sex Roles*, **21**, 803-823 (1989).
24. C. Tausanovitch, C. Warshaw, Representation in municipal government. *Am. Polit. Sci. Rev.* **108**, 605-641 (2014).
25. U. S. Census Bureau, American Community Survey, 2015-2016
26. P. T. Cohen-Kettenis, M. Wallien, L. L. Johnson, A. F. H. Owen-Anderson, S. J. Bradley, K. J. Zucker, A parent-report gender identity questionnaire for children: A cross-national, cross-clinic comparative analysis. *Clin. Child Psychol. Psychiatry*. **113**, 397-405 (2006).
27. M. S. C. Wallien, L. C. Quilty, T. D. Steensma, D. Singh, S. L. Lambert, A. Leroux, A. Owen-Anderson, S. J. Kibblewhite, S. J. Bradley, P. T. Cohen-Kettenis, K. J. Zucker, Cross-national replication of the gender identity interview for children. *J. Pers. Assess.* **91**, 545-552 (2009).
28. C. Leaper, "Gender and Social-Cognitive Development" in *Handbook of Child Psychology and Developmental Science*, R. M. Lerner, Ed. (Wiley, 2015).
29. D. Lakens, Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Soc. Psych. Pers. Sci.* **8**, 355-362 (2017).