

Supplementary Material

This supplement includes (1) additional methodological details, (2) an additional table for the comparison amongst CAH girls and girls without CAH, (3) additional information about the toy and clothing measure (including piloting details), and (4) information about a parental measure of children’s gendered behavior and associated results.

Additional Methodological Details

Families were told about the possibility to complete the study before, after, or between their visits with medical care providers at one of the following types of clinics: differences in sex development (DSD) clinic, pediatric endocrinology clinic, CAH clinic, or Klinefelter clinic. To be included a child must have been judged by medical providers as having a VSD, the child had to be between 3 and 12 years of age and have enough English comprehension and production to complete the study. Children with severe intellectual impairments who would be unable to answer these questions and children who did not speak English were excluded from recruitment. Nearly all participants completed the research session during or immediately after a clinic visit, but one did so on Zoom later.

There are two pairs of siblings among the 78 participants with VSDs and one pair of siblings among the community comparison participants; these six individuals are treated independently for these analyses.

Participation for community controls occurred either in person at a research laboratory (n=39) or online through Zoom (n=39). Both university databases exist for the recruitment of child development research participants. Families often sign up to be included in the database when the children are born or later at a community event. They are told that researchers studying a range of topics recruit from these databases (e.g., language development, prosocial behavior, infant cognition). Parents were told they could make decisions later about which studies they would or would not participate in. All parents of these community participants were told that this was a study about gender development and agreed to have their child participate.

Assent involved hearing a short script describing the study (using different, developmentally appropriate language for 3-8-year-olds vs. 9-12-year-olds) indicating that participants could choose to participate or not, skip questions, and stop participation at any time. The child answered questions one-on-one with a researcher on a tablet or laptop, while parents completed a questionnaire on paper or on a tablet (typically in another room; two parents did not complete the questionnaires).

Toy and Clothing Preference Measures-Further Information

There were different stimuli for children aged 3-7 years old and for children aged 8-12 years old because children at these ages play with different toys and wear different outfits. In addition, thirty-two youth (16 youth with VSDs and their 16 community comparison youth) completed these tasks with the stimuli from existing publications (7,31) that had originally been pilot-tested in 2013-2014. One hundred and twenty-four youth (62 youth with VSDs and their 62 community comparison youth) completed these tasks with the same structure, but with updated toy and clothing stimuli (updated as described below). We treat these two sets of stimuli as equivalent for analytic purposes; results hold when analyzing the results of only the 124 youth receiving the second set of stimuli.

**Pilot testing of new items:** Pilot testing of the exact stimuli used as part of the revised toy and clothing measure involved the following steps. First, children between the ages of 3 and 12 (who were not part of the main study) were shown outfits or toys thought by study staff to be appropriate for their age and they were asked “Who would play with this toy” or “Who would wear this outfit?”. Each child saw 25 clothing and 25 toy options, selected from a larger set of 100 total items of each type. They responded to each of the items with one of 5 selections: only boys, mostly boys, both boys and girls, mostly girls, and only girls. These were converted to scores from 1 (only boys) to 5 (only girls). The average score was then computed for each item. Using these scores, we created 4 sets of 5 toys that were considered strongly feminine, moderately feminine, neutral, moderately masculine, and strongly masculine.

In the second round of piloting, a new group of 21 7-11 year old children (also participants who were not in the main study) saw a set of 5 items that included one each of each rating (e.g., strongly feminine, moderately feminine, etc.) and were asked to select one item from each set to place into each of five categories: many more girls than boys wear/play with the item, slightly more girls than boys wear, play with the item, boys and girls equally wear/play with the item, slightly more boys than girls wear/play with the item, many more boys than girls wear/play with the item. Children were asked to answer based on what most kids would think. We used this second round of piloting to confirm that the selections from the first part of piloting were valid in representing these categories to a new group of participants. Further, we wanted to ensure that the sets of 5 items we placed together into a set did not somehow impact the relative perception of their masculinity/femininity. We found that children sorted the items as expected based on the initial ratings and therefore used these new item sets for the remainder of the study participants. We did not ask younger children to complete this secondary part of the piloting process because it seemed too challenging, but we did ask the 7-11 year olds to rate both the younger and older stimuli, assuming their opinions might be closer to those of younger children than adult raters (though research staff also confirmed that the sets seemed face valid as well).

Parent-Reported Youth Gender Identity and Experience

While not the focus of the primary analyses for the current paper, parents completed a measure of their perception of children’s gender identity and expression. We provide that information here so as not to distract from the focus on children’s own gendered behavior and preferences in the main paper.

We used two similar questionnaires asking parents about their children’s gender-typed preferences, behaviors, femininity/masculinity, interests, and feelings about their body, etc. Each item was scored from 0-1 where 1 indicates either the response most stereotypically aligned with the gender in which the child was raised, or the child’s highest possible satisfaction with their gender and body (0 represents the response least stereotypically aligned with the child’s gender-of-raising, or the child’s lowest possible satisfaction with their gender and body). Responses indicating that the child does not participate in an activity of a particular kind are omitted. We use the average score across items and treat the two scales as equivalent for research purposes ( $\alpha_{\text{RCGIE-GS}}=0.85$ ;  $\alpha_{\text{GIQ}}=0.66$ ).

Most parents answered the Recalled Childhood Gender Identity and Experience-Gender Spectrum-Parent Version (RCGIE-GS), a 32-item measure (47) adapted from (48) (we further adapted this measure to ask parents about their youth in the present tense; 62 parents of youth with a VSD were invited to complete this measure). A smaller number of parents completed an 18-item version of the parent-report Gender Identity Questionnaire for Children (GIQ; (49); we adapted this measure so all parents answered two boy-specific and two girl-specific items; 16 parents of children with VSDs were invited to complete this measure). We switched from (49) to (47) because it uses more contemporary language; however, questionnaire version was matched between youth with VSDs and their community control youth.

Results involving the RCGIE-GS/GIQ

Parents of youth with VSDs and community comparison parents did not differ in their ratings of their child’s gender typicality on the RCGIE-GS/GIQ [t(145.31)=0.764, p=0.446, d=0.124;  $N_{\text{VSD}}=75$ ,  $M_{\text{VSD}}=0.75$ ,  $SD_{\text{VSD}}=0.11$ ;  $N_{\text{Community}}=77$ ,  $M_{\text{Community}}=0.76$ ,  $SD_{\text{Community}}=0.09$ ].

To evaluate the relationship between parent and child report, we reverse-scored the child-report composite for children raised as boys, so that higher scores represented greater gender typicality in child behavior in the experimental tasks. Parent reports of children’s gendered behavior on the RCGIE/GIQ were positively correlated with the gender-typicality composite [r(150)=0.51, p<0.001].

| Table S1. Within-group means for measures of gendered self-report between CAH girls and comparison participants                                                                         |                   |             |                      |             |                                        |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|-------------|----------------------|-------------|----------------------------------------|
|                                                                                                                                                                                         | Children with VSD |             | Community comparison |             | Raised-as girls                        |
| Measure                                                                                                                                                                                 | N                 | M (SD)      | N                    | M (SD)      | t-test                                 |
| Toy preference                                                                                                                                                                          | 28                | 0.44 (0.31) | 28                   | 0.75 (0.21) | t(48.22)=4.314<br>p<0.001<br>d=1.153   |
| Clothing preference                                                                                                                                                                     | 28                | 0.76 (0.20) | 28                   | 0.82 (0.19) | t(52.76)=1.176<br>p=0.245<br>d=0.314   |
| Peer preference                                                                                                                                                                         | 28                | 0.78 (0.20) | 28                   | 0.76 (0.22) | t(53.09)=-0.506<br>p=0.615<br>d=-0.135 |
| Continuous gender identity                                                                                                                                                              | 28                | 0.92 (0.14) | 28                   | 0.89 (0.18) | t(51.71)=-0.687<br>p=0.495<br>d=-0.184 |
| Similarity to boys                                                                                                                                                                      | 22                | 0.65 (0.27) | 27                   | 0.74 (0.24) | t(42.45)=1.327<br>p=0.192<br>d=0.383   |
| Similarity to girls                                                                                                                                                                     | 22                | 0.69 (0.23) | 28                   | 0.68 (0.31) | t(47.76)=-0.215<br>p=0.831<br>d=-0.060 |
| Each measure ranges from 0 (most stereotypically associated with boys) to 1 (most stereotypically associated with girls).<br>VSD: variations in sex development; SD: standard deviation |                   |             |                      |             |                                        |