Published in final edited form as:

Soc Dev. 2022 August ; 31(3): 916–929. doi:10.1111/sode.12587.

Between a boy and a girl: Measuring gender identity on a continuum

Selin Gülgöz[1,2,*], **Deja L. Edwards**[2,*], **Kristina R. Olson**[2]

[1]Fordham University, New York, NY

[2]University of Washington, Seattle, WA

Abstract

Studies of gender development typically use binary, discrete measures of identity. However, growing literature indicates that some children might not identify with a binary gender. We tested a continuous measure of felt gender identity with gender-nonconforming children, socially-transitioned transgender children, cisgender siblings, and unrelated cisgender children. In two studies, we found that transgender and cisgender children did not differ in their degree of identifying as their current gender, that they showed more binary identities compared to gender-nonconforming children, and that the continuum was a valid predictor of other measures of gender development. We also found that children's use of the continuous measure was stable over time. Our results demonstrate the test-retest reliability and validity of a new single-item continuous measure of gender.

Keywords

gender identity; transgender; gender diversity; gender development

With a few notable exceptions, most studies that assess gender identity use measures with two discrete, binary choices: "Are you a boy or a girl?" This measure can be used with children as young as two or three years old, which is when most children can label their own genders using terms like "boy" and "girl" and use these labels to group themselves and others into separate gender categories (e.g., Etaugh et al., 1989; Fagot et al., 1992; Weinraub et al., 1984; Zosuls et al, 2009). There are clear advantages of simple measures when conducting research with children. However, there is growing evidence that at least by adolescence, some people do not think of themselves in discrete, binary ways, rather use terms like *nonbinary* or *genderqueer* (Diamond, 2020; Hyde, et al., 2019; Salk et al., 2020) that are not captured when binary measures are adopted. Thus, discrete or binary measures

Corresponding author: Selin Gülgöz; sgulgoz@fordham.edu.
*Co-first authors

Conflicts of interest/Competing interests. Kristina Olson has an (unpaid) position on the Big Brothers Big Sisters of America LGBTQ National Advisory Council and receives compensation from the MacArthur Foundation. Though we do not perceive these as conflicts of interest, we declare them in the spirit of full transparency.

Availability of data and material. De-identified data and materials used are stored on OSF and can be accessed at this link: https://osf.io/9tsr5/.

Code availability. The R script used to analyze the data is available on OSF: https://osf.io/9tsr5/.

like the one described above prevent the ability of researchers to investigate nuance or variation in gender identity.

Less discrete measures of gender identity in children

There is no clear consensus in the field on the definition of gender identity and how it should be measured (Mehta, 2015; Tobin et al., 2010). Some have measured gender identity by focusing on children's perceptions of how similar they are to members of one gender group or another (e.g., Egan & Perry, 2001; Martin et al., 2017), whereas others have focused on how a child feels on the inside or a child's felt sense of their gender (e.g., Gülgöz et al., 2019; Reisner et al. 2015; also the definition used in the current paper). A third approach has been to conceptualize gender identity as a multi-faceted construct, requiring multiple types of measures used in combination (e.g., Egan & Perry, 2001; Dinella et al., 2014). Within these different perspectives, there have been a few studies that measure gender identity in children using more continuous methods. Egan and Perry (2001) developed a multidimensional approach, asking children to select which of a pair of statements was truer of them, e.g., "some girls think they are a good example of being a girl" vs. "other girls don't think that they are a good example of being a girl," and to what degree. Though useful in assessing variation in children's gender identities, this 92-item measure requires considerable linguistic aptitude and stamina, and is therefore used with children in later elementary and middle school, and not necessarily younger children (though note that the Gender Typicality subscale, which is perhaps most relevant to the current paper, is six items long). Martin and colleagues (2017) built upon the work of Egan and Perry (2001), with a simplified task for use with elementary-aged children. Children were asked 10 questions like "How much do you look like girls?" and "How similar do you feel to boys?" and replied using one of five pictorial replies indicating a degree of similarity.

There have been considerable benefits to these approaches. Responses on these measures have been linked to meaningful real-world outcomes such as social adjustment, acceptance by peers, self-esteem, and mental health (e.g., Martin et al, 2017; Yunger et al., 2004). These measures have also demonstrated that simply "boy" or "girl" does not capture the range of experiences of many children. For example, Martin et al. (2017) discovered that some subsets of children feel strongly connected to both genders, whereas others feel more connected to their own gender than the other gender.

At the same time, these past measures are designed for older children and involve many questions with complex scales. Both measures focus on children's perceptions of their own gender typicality (among other constructs like gender contentment, felt pressure to conform, etc. in Egan and Perry's measure), rather than a direct assessment of children's self-labeling into a felt gender category. A simplified variation of the Martin et al. (2017) similarity measure was used in an earlier study (Zosuls et al., 2016), where researchers found that five-year-old children were able to reliably use a two-item version of the similarity task (i.e., asking participants the two questions, "How much are you like girls?" and "How much are you like boys?"). However, both the Egan and Perry (2001) and Martin et al. (2017) measures relied on children's social projection skills (i.e., comparing themselves to girls and boys as social groups) rather than directly asking about a child's gender identity the way

traditional tasks have done. The more traditional and simple measure of "Are you a boy or girl?" has the advantage that it involves only one question, is not comparative, requires little to no complex language or scales, and can be used with preschoolers. As such, it can be easily used with only a minimal increase in testing burden, even in studies that are not about gender.

The current work

In the present work, we investigated a new single-item, continuous measure with simplistic language appropriate for use with children as young as three years of age. Although we test the measure through age 14, we expect it could be used with older adolescents and adults as well). The goal of this measure is to complement existing measures of gender identity, and use it in research that might not be directly about gender but might seek to incorporate a non-categorical measure of gender identity.

We tested the measure using several assessments of reliability and validity. In Studies 1a and 1b, we tested whether responses on the measure were correlated with other conceptually related measures—the Martin et al (2017) measure, measures of gender-typed preferences (Fast & Olson, 2018) and a measure of implicit gender identity (Olson et al, 2015). Previous work has found that the Martin et al. (2017) measure of gender identity, when scored as a difference of "similarity to own gender" minus "similarity to other gender", is correlated with the same gender-typed preferences measure (Gülgöz et al., 2019). In Study 2, we examined test-retest reliability over a period of one to two years.

The present paper included children usually tested in such studies—cisgender children (i.e., their gender identity and expressions tend to align with their assigned sex at birth)—but also children who are gender diverse. We tested our measure in a group of socially-transitioned binary transgender children—i.e., they use "he" or "she" pronouns but this gender does not align with their assigned sex—and a group of gender-nonconforming children—children who defy gender norms for their sex at birth, but either continue to use their original pronouns or have less binary identities (e.g., use "they" or similar nonbinary pronouns). Past work has suggested these two groups differ on measures of gender identity and preferences (e.g., Rae et al, 2019), but that transgender and cisgender groups often do not differ (Gülgöz et al., 2019), allowing us to test if the same patterns emerge on this measure as well.

Prior to data analysis, but after some data had been collected, we registered our samples, research questions, and analytic approach for Study 1a on OSF (https://osf.io/d7eau). According to this registration, we were going to recruit transgender and gender-nonconforming participants until we reached at least 40 gender-nonconforming children, or until October 1, 2017, whichever occurred first. After completing that study, however, we continued data collection (for Study 1b, a replication of Study 1a) until January 10, 2020, after which in-person data collection with gender diverse youth was halted due to COVID-19. In Study 2, we assessed test-retest reliability, including all participants who had completed the measure at least twice before COVID-19 ended in-person data collection. The current studies were approved by the University of Washington Institutional Review Board (approval #00001527).

Studies 1a and 1b

Method

Participants—Participants in the current studies were part of larger, often longitudinal studies on gender development (Fast & Olson, 2018; Gülgöz et al., 2019; Olson et al., 2015; Rae et al., 2019; Rae & Olson, 2018). Participants belonged to one of four groups: (1) transgender, (2) gender-nonconforming, (3) cisgender siblings of transgender/gender-nonconforming, and (4) unrelated cisgender children. Demographic information for each participant group is presented in the online Supporting Information (Table S1). There was no overlap in participants of Study 1a and Study 1b.

Transgender Participants.: Transgender participants were recruited in the U.S. and Canada through conferences and camps for gender-diverse children, media coverage, word-of-mouth, and support groups, or their parents signed up on our website and testing occurred in participants' homes or local settings. At the time of testing, all transgender participants had socially transitioned, which we had operationalized as using pronouns that are typically associated with the binary gender different from the one they were assumed to have at birth (i.e., an assigned female who uses "he" pronouns in all social contexts would be a trans boy in this study).

Study 1a included 132 transgender children (M_{age} = 8.44 years, 3.50-14.08 years, SD = 2.40 years; 85 transgender girls/assigned males, 47 transgender boys/assigned females). Study 1b, included the next 168 transgender participants (M_{age} = 8.76 years, 4.17-13.67 years, SD = 2.11 years, 121 transgender girls/assigned males, 47 transgender boys/assigned females).

Gender-Nonconforming Participants.: Gender-nonconforming participants were defined as children who, according to their parents, showed identity, preferences and/or behaviors that are stereotypically associated with a gender different from that typically associated with the sex they were assigned at birth. Different from our transgender group, gender-nonconforming participants had not socially transitioned (i.e., were not using binary pronouns associated with a gender different from the one they were assigned at birth) at the time of study. Examples of gender-nonconforming children in this study would be an assigned male who uses "he/him" or "they" pronouns and is described by their parent as showing interest in stereotypically feminine and/or gender-neutral clothing (e.g., dresses), toys (e.g., Barbie dolls) or media characters (e.g., Disney princesses), etc. A previous report with this sample of gender-nonconforming youth (Rae et al., 2019) indicates that parent-report did align with youth's actual preferences, which tended to align more with cisgender youth of the other binary sex at birth than youth of the same sex at birth.

Gender-nonconforming participants were recruited using the same methods as described above. Study 1a included 38 gender-nonconforming children (M_{age} = 7.74 years, 3.25-14.08 years, SD = 2.97 years; 14 assigned females, 24 assigned males). Study 1b included 55 gender-nonconforming participants (M_{age} = 7.98 years, 3.33-12.67 years, SD = 2.44 years, 19 assigned females, 36 assigned males).

Siblings.: Cisgender siblings of transgender or gender-nonconforming participant in the study were also included as participants. Siblings were recruited if they were between three and 12 years old. If there was more than one sibling, the sibling closest in age was recruited, unless that sibling was unavailable or had another reason not to participate (e.g., a developmental delay). In Study 1a, this group included 104 siblings (M_{age} = 8.27 years, 3.00-14.41 years, SD = 2.80 years; 46 girls, 58 boys; 78 siblings of transgender participants; 26 siblings of gender-nonconforming participants). In Study 1b, there were 138 cisgender siblings (M_{age} = 8.41 years, 3.67-14.58 years, SD = 2.42 years, 64 girls, 74 boys; 107 siblings of transgender participants; 31 siblings of gender-nonconforming participants).

Age- and Gender-Matched Cisgender children.: We recruited cisgender participants who were not related to the transgender or gender-nonconforming participants. The children in this group were matched by age to each transgender and gender-nonconforming participant and had the binary gender that differed from the transgender and gender-nonconforming participants' assigned sex (based on the matching method used in previous work, Glazier et al., 2020; Olson et al., 2015). Cisgender children were recruited via a participant database through the University of Washington, and testing was conducted locally in the lab. In Study 1a, there were 165 unrelated cisgender children (M_{age} = 8.64 years, 3.25-14.25 years, SD = 2.20 years, 152 girls, 62 boys). In Study 1b, unrelated cisgender children included 214 children (M_{age} = 8.64 years, 3.16-13.83 years, SD = 2.20 years, 152 girls, 62 boys).

Measures and Procedure—The measures and procedure for Studies 1a and 1b were identical.

Gender Spectrum.: We used a continuum to assess gender identity (called the "Gender Spectrum" task; see Figure 1). To measure a child's sense of their own gender identity, we showed participants a continuum ranging from "feeling totally like a boy" to "feeling totally like a girl" and asked them to "…put an X in the place you think best shows how you *feel* on the inside". This wording was selected to align with how researchers often conceptualize gender identity (e.g., Gülgöz et al., 2019; Reisner et al., 2015) as well as how gender is often discussed amongst transgender people (e.g., not necessarily as what is associated with their sex assigned at birth, but the gender they feel they are, Olson, 2017). Participants' responses were scored by measuring the distance between the left end of the line and the 'X' they placed on the line and converting this score to a percentage of the whole line. This was measured by two independent scorers, and high inter-rater reliability was achieved (Studies 1a and 1b, α = 0.99 and 0.99, respectively). There were a few cases where the difference between the two coders' scores was greater than 10 (Study 1a: seven participants; Study 1b: eight participants). Because this was a physical measurement, in these cases, we knew (at least) one coder was objectively incorrect. Therefore, a third independent coder also measured the score in these cases, and we selected the two scores that were closer to each other as the accurate scores. For each participant, the two final coder scores were averaged to produce a single final score.

Scores ranged from 'feeling totally like a boy' (0) to 'feeling totally like a girl' (100). Following scoring standards used with similar tasks and utilizing similarly gender-diverse samples in prior research (Rae et al., 2019), after initial scoring, scores were reversed

for transgender boys, cisgender boys, and gender-nonconforming girls so that low scores aligned with assigned sex at birth, and higher scores aligned with current felt gender for transgender and cisgender participants. In other words, for both transgender and gender-nonconforming participants, higher scores indicated identifying as or associating more closely with a gender that is different from their sex assigned at birth, as previous research has suggested that this is consistent with the general patterns of identity and gender typing seen for many gender-diverse youth (Rae et al., 2019). This same scoring method was also used for all measures described below.

Because this study was part of a larger study, the participants in the current study also completed several other measures related to gender development (e.g., essentialism) and development in general (e.g., mental health), but this report focuses on the measures we registered before data analysis, with one exception noted below. To test the criterion validity of the continuous gender identity measure, we examined the extent to which scores on the continuum correlated with three other sets of measures relevant to gender development.

Toy, Clothing and Peer Preferences.: In the preferences measure adapted from Fast & Olson (2018), we assessed the extent to which children showed preferences for stereotypically girl- or boy-typed toys, clothing, and peers. For the toys and clothing preference tasks, participants were shown four arrays of toys and four arrays of clothing. Each array included five toys or outfits that had been previously pretested to range from highly stereotypically masculine to gender-neutral to highly stereotypically feminine. In each array, children were asked to select which toy/outfit they would like to play with/wear the most. In the peer preference task, participants were shown six pairs of children (each pair consisted of one girl and one boy) and were asked to select who they would like to be friends with. A combined score was calculated for all three tasks and participants received a score out of 100, with lower scores indicating preferences associated with their sex assigned at birth.

Similarity.: Additionally, we used Martin and colleagues' (2017) similarity measure described above, to assess participants' perceptions of their similarity to members of their own and other gender. Children were asked a total of 10 questions, half of the questions assessing similarity to girls; half of the questions assessing similarity to boys. These questions included items like "How much do you look like girls/boys?" and "How similar do you feel to girls/boys?". Participants replied using a five-point pictorial scale in which they saw two circles of increasingly close distance (through overlap). One of the circles was smaller and represented the child participant, and the other circle was larger and represented the question the group was about (e.g., boys). Borrowing the approach used in Gülgöz et al. (2019), we separately measured each participant's similarity to girls and similarity to boys, and subtracted the two scores from each other.

Gender Identity Implicit Association Test.: We borrowed the version of the Implicit Association Test (IAT) used in Olson et al. (2015), designed for understanding children's implicit identification with boys vs. girls. For analyses, an IAT D score was computed using a standard algorithm (Greenwald, Nosek, & Banaji, 2003). Higher scores indicated higher implicit identification with girls, and lower negative scores indicated higher implicit

identification with boys; these scores were converted according to the principles outlined above. Across participant groups, children who made errors on more than 30% of trials (Study 1a: $n = 0$; Study 1b: $n = 11$), or who completed more than 10% of their responses in less than 300 ms (a speed that is too fast to process stimuli) (Study 1a: $n = 2$; Study 1b: $n = 0$) were excluded. Because this test required participants to read, children under six years of age or who were not able to read were not tested on this measure so the sample size for the analyses involving the IAT are always smaller.

Categorical Gender Identity.: Finally, as a post-hoc exploration, we examined the extent to which participants' responses on the new continuum measure corresponded to the responses they gave to a more categorical measure we used. In this categorical gender identity measure, participants were verbally asked "Do you feel like you are a boy, a girl, or something else?" If a participant responded with the answer "something else," they were given additional options to choose from: "both," "neither," "it changes," and "I don't know."

The preference, similarity, and IAT scores, and responses to the categorical gender identity measure of the current studies' participants and others participating in the larger longitudinal project have been previously reported in Gülgöz et al. (2019), Rae et al. (2019).

Results

As registered, in Studies 1a and 1b, we conducted one-way ANOVAs examining the effect of participant group (4: transgender, gender-nonconforming, unrelated cisgender children, cisgender siblings) on spectrum scores. We found a significant main effect of participant group in Study 1a, $F(3,427) = 21.70$, $p < .001$, $\eta_p^2 = .13$, and in Study 1b, $F(3,558) = 24.60$, $p < .001$, $\eta_p^2 = .12$. Post-hoc Tukey comparisons showed that both in Study 1a and Study 1b, transgender participants, unrelated cisgender children and cisgender siblings did not differ from each other in their degree of identifying with their own gender (ps ≥ .466); gender-nonconforming participants significantly differed from all other groups (ps < .001; see Table 1 for descriptive statistics and Figure 2 for histograms).

We used one-sample t-tests to compare each participant group's score to the midpoint of the scale (.50; identifying equally as boy and girl) and ceiling (1; identifying totally as their own gender). As seen in Table 1, participants in all groups scored significantly above the midpoint, with the exception of gender-nonconforming children, who were not different from the midpoint in Study 1a; all groups were significantly below the ceiling in both studies.

We next assessed the extent to which the spectrum measure correlated with other measures of gender identity and gender typing, including preference scores, similarity scores, and IAT scores, within each participant group (see Table 2; also see Figure S1 in the online supplement for exploratory analyses).

As registered, we only conducted correlation analyses with transgender participants, unrelated cisgender children and siblings, as the number of gender-nonconforming participants who completed each task were too low to reliably conduct correlation analyses with (ranging from 15 to 30 across tasks). Across both studies, we found that scores

on the spectrum measure were significantly positively correlated with preference scores for unrelated cisgender children and siblings, indicating that those who identified more strongly with their own gender on the spectrum measure also showed more gender-typical preferences in toys, clothing and peers. Spectrum scores were also significantly positively correlated with similarity scores for all groups, meaning that higher identification with own gender on the spectrum was related to greater perceived similarity with own gender as well (see Table 3). With the exception of the unrelated cisgender children in Study 1a (a small and negative effect), scores on the IAT were not significantly correlated with spectrum scores for any participant groups.

Among transgender participants, we examined the extent to which time since transition predicted participants' responses on the spectrum measure (registered as a secondary research question). Because time since transition was correlated with participant age (Study 1a: $r(130) = .86$, $p < .001$; Study 1b: $r(165) = .68$, $p < .001$), we conducted partial correlations controlling for participant age, and found that transgender participants' gender identities, as measured on the spectrum, did not differ by how long they had lived as their current gender (Study 1a: $r(129) = -.01$, $p = .922$; Study 1b: $r(165) = .06$, $p = .438$).

Discussion

In Studies 1a and 1b, we found that transgender children, their cisgender siblings and unrelated cisgender children did not differ from each other in the extent to which they identified with their current gender on a new continuous gender identity measure. All three groups clearly identified with their current gender more than the other gender. This means that a transgender girl and a cisgender girl were equally likely to identify as a girl even when given the chance to express their identities on a spectrum. In contrast, gender-nonconforming children showed greater variability across the scale and their average responses differed from all three other groups.

These results support previous findings with categorical measures of gender identity (Gülgöz et al., 2019). However, these findings also refine what previous studies demonstrated. None of the groups in Studies 1a and 1b identified exclusively as their own gender (as would be indicated by an average score at the end point), even if some children within each group did. That is, even among cisgender and transgender participants, not all children gave the most extreme responses on the spectrum measure. This finding suggests that children's gender identities cannot be characterized as completely binary.

Scores on the spectrum measure generally were associated with scores on Martin and colleagues' (2017) similarity measure. An advantage of our measure is that it involved a single item and is quick to complete, even for young children, whereas an advantage of the Martin et al. (2017) 10-item measure or the Zosuls et al. (2016) two-item measure is that they have multiple items that create a reliable scale. These prior measures also have the advantage that they might show children's perceived similarity to boys and girls independently of each other, whereas the new spectrum measure forces participants to choose a position on a continuum where boy and girl are presented as relative to each other. Another difference between the two types of measures is that the current measure assesses the extent to which children viewed themselves as a girl or a boy or some combination,

whereas the prior measures relied on children's perceptions of themselves in comparison to others. Thus, there may be cases where one measure versus the other might be more ideal for a research study (e.g., depending on the research question, the age of participants or how much time there is for a measure).

For cisgender participants, the spectrum measure was also correlated with preferences for items like toys and clothing that are gendered in society. In contrast, the spectrum measure did not predict gender-typed preferences for transgender children in either study. This is surprising because the mean score on the spectrum and preferences measures for the transgender group did not differ from the cisgender groups in either study. Additionally, we found that scores on the similarity measure were positively correlated with preference scores for transgender (as well as cisgender) participants in both studies (ps < .001). Thus, it is difficult to know why there is not a significant relation between the spectrum and preference scores of the transgender children in this sample (though note that it is also the case that the correlation for the transgender group did not differ in magnitude from the correlation for the unrelated cisgender group despite one being significant and the other not; Fisher r-to-z transformations: Study 1a: $z = .73$, $p = .465$; Study 1b: $z = .84$, $p = .401$). Further testing of the predictive power of the spectrum measure, as well as correlations between identity and gender typing measures among gender-diverse samples are needed to develop a clearer conclusion.

Finally, we did not find significant correlations between spectrum scores and IAT scores. This is in line with previously published findings showing that correlations between the IAT and other measures of gender identity vary considerably depending on whether one scores the IAT according to participants' own gender vs. other gender (as we have done in this study) or scoring it as boy vs. girl. The reason for this is a variant of the Simpson's paradox. The IAT appears to be good at sorting children, especially among binary-identified children, into two clusters that represent the categories boys and girls. Within each cluster, there is little or no positive correlation between the degree of identification with one's gender and any of the other measures (something we also see when coding the IAT as own vs. other gender). However, because there are two clusters, when scoring responses from boy to girl, a positive relationship emerges (just as a line can be made from any two points, so too, for any two clusters; for more detail on this issue, see Gülgöz et al., 2019; Rae & Olson, 2018).

Study 2

To assess the reliability of this new measure over time, we examined whether there was a relation between children's response on the spectrum measure when assessed twice at least 10 months apart. We included all children who were given this measure at least twice in the course of the larger longitudinal project before our lab closed for in person testing due to COVID-19.

Participants

Participants included a subsample from Studies 1a and 1b of 196 children who had answered the spectrum measure in at least two separate testing sessions. Testing sessions were 1.95 years apart on average (range 0.85 to 4.57 years). The final sample included 77 transgender

participants (M_{age} = 9.03 years, SD = 1.60; 55 transgender girls, 22 transgender boys), 14 gender-nonconforming participants (M_{age} = 8.58 years, SD = 1.72; 6 assigned females, 8 assigned males), 41 sibling participants (M_{age} = 6.97 years, SD = 2.28; 22 girls, 19 boys), and 64 unrelated cisgender participants (M_{age} = 6.96 years, SD = 1.57; 45 girls, 19 boys).

Results and discussion

We conducted two sets of preliminary analyses. First, we conducted an independent samples t-test to compare spectrum scores of participants from Studies 1a and 1b who were (n = 196) and were not (n = 818) included in Study 2. We found that the two samples did not significantly differ from each other in their Time 1 spectrum scores, $t(991) = .41$, $p = .679$. Second, to see whether findings from Studies 1a and 1b replicated when participants responded to the spectrum measure during their second visit, we conducted a one-way analysis of variance comparing the four groups' responses at Time 2. We found a significant effect of participant group on spectrum scores, $F(3,188) = 9.86$, $p < .001$, $\eta_p^2 = .14$. Like in Studies 1a and 1b, the transgender participants, cisgender siblings and unrelated cisgender children did not differ from each other, but gender-nonconforming children's scores were significantly less binary when compared to transgender and cisgender participants (ps < .002, post-hoc Tukey tests; see Table 4 for descriptive statistics).

Due to the small sample sizes in each of the groups, we collapsed across participant groups for our primary analysis. Scores on the spectrum measure were positively correlated across a two-year testing gap, $r(181) = .36$, $p < .001$. This relation was not affected when we controlled for age at first visit, $r(183) = .34$, $p < .001$, nor when we controlled for time passed in between visits, $r(183) = .36$, $p < .001$. We also calculated the amount of change each participant had shown on the spectrum from one visit to the next. We found that the absolute value of the difference between participants' two scores ($M_{difference}$ = 1.17) also did not correlate with the time in between visits, $r(181) = .11$, $p = .145$. Thus, in general, participants' gender identity responses on the spectrum measure were moderately stable across time and development (mean results by group are listed in Table 4; also see the online Supporting Information, Figure S2, for individual patterns).

Though previous research on stability of gender identity is limited, to contextualize these findings better, we can look at the few studies that have examined the extent to which children's early gender identification and gender-typed preferences relate to their later gender identity and gender-typed preferences. Rae and Olson (2018) found that gender-diverse children's scores on the IAT were highly correlated across a year (r = .56). Similarly, DeLay and colleagues (2018) assessed 6th graders' perceptions of similarity to members of their own gender at the beginning and end of a semester, and found high stability (r = .63). Examining the stability of same-gender peer preferences among seven and nine-year-olds, Halim and colleagues (2018) found moderate correlations across a 12-month period (r =.33).

Although it is difficult to compare these past findings to the current one, due to differences in variables like measures used, participant groups, and ages, to further explore the stability rate found in the current study, we conducted additional correlations examining stability among transgender and cisgender (i.e., siblings and unrelated cisgender children) participants. Our reason for conducting these post-hoc analyses was a suggestion from

anonymous reviewers that the observation that the mean difference in scores for siblings appeared to be higher than the mean differences in scores for other groups (see Table 4). We found the rates of stability for these groups to be $r = .27$ and $r = .26$, respectively (both significant at $p \leq .02$). It's important to note that these exploratory analyses have significantly lowered statistical power due to smaller sample sizes, and should therefore be interpreted with caution.

General Discussion

Consistent with prior research using categorical measures of gender identity, the current work shows that transgender children, cisgender siblings and unrelated cisgender children do not differ in the extent to which they identify as their current gender on a spectrum measure of gender identity (Gülgöz et al., 2019). Moreover, neither cisgender nor transgender children identified exclusively as their current gender, which suggests that continuous measures might provide more nuanced representations of children's gender identities. These results lend support to the growing recognition that binary gender categories might not fully reflect how individuals identify (Diamond, 2020; Morgenroth & Ryan, 2020).

Though they comprised a much smaller sample, the gender-nonconforming children in Studies 1a and 1b were a more heterogeneous group in their use of the spectrum measure, compared to the transgender and cisgender groups. Whereas transgender and cisgender children, on average, identified closer to the end of the spectrum, gender-nonconforming children were more spread out with a resulting average in the midpoint of the spectrum. Previous research with this group has shown that there might be distinct developmental pathways within this group (Rae et al, 2019) and that they tend to show greater variability on several measures of gender development. Our findings are consistent with this previous work. Our findings are also consistent with previous theorizing by Tate and colleagues (2014) in terms of showing the similarities between transgender and cisgender individuals' gender identities, where neither group shows fully binary identification, and in terms of showing that there are individuals who show different, less binary gender identities.

One possible caveat concerning the findings from the gender-nonconforming group has to do with how their responses were scored. Following previous scoring standards (Rae et al., 2019), we scored each measure so that higher scores would indicate closer identification or association with the gender that is different from a gender-nonconforming participant's sex assigned at birth. The main reason for this was because previous research (Rae et al., 2019) has suggested this to be the general direction for most youth in this group. However, this scoring may not have appropriately represented gender identities of youth who identify as nonbinary (e.g., neither boy nor girl, or both boy and girl). Inclusion of nonbinary youth might explain the greater variability of scores found among the gender-nonconforming group.

Studies 1a, 1b, and 2 demonstrate initial validity of this spectrum measure of gender identity. This was evident from findings suggesting that participants' scores on the spectrum measure can predict scores on another gender identity (Martin and colleagues' similarity measure) and that there is test-retest reliability, even an average of two years later. The spectrum

measure was also correlated with gender-typing measures (toy, clothing, peer preferences) in our cisgender samples. Comparisons to participants' responses on a categorical measure of gender identity (see Supporting Information) also demonstrated the extent of possible variation of individuals' felt gender identities even when they are grouped into the same binary category. Even for the youngest participants (i.e., three- to five-year-olds), the spectrum measure appeared to be easy to understand and use, and representative of gender identities that are typically measured with categorical measures (see Supporting Information). However, although we designed this measure to be used with a broad range of age groups, including children as young as preschool-aged, a larger study of younger children would be necessary to see if it really is valid in this three- to five-year age group.

Findings from the current work raise several methodological and conceptual questions for future work. First, an important limitation to consider in the current work is that the children tested in these studies come from families who are highly supportive of their child's gender identity expressions, and who are predominantly White, of high socioeconomic status, and liberal. Therefore, the extent to which the findings in this paper are representative of children from other backgrounds within the U.S. and across the world is unknown and remains an open question for future research. Second, although the spectrum measure provides a continuum for children to identify their gender on, it is still situated within the gender binary by virtue of having two poles and placing them opposite each other, labeled with terms "boy" and "girl". Thus, although the measure allows for assessing and demonstrating greater nuance of gender identities within the binary, it may not accurately capture nor distinguish, for example different nonbinary, fluid, agender, or other gender identities. It would be important to test whether other measures that utilize multiple spectrums or questions (e.g., Martin et al., 2017; Zosuls et al., 2016, or new measures to be developed) might better represent nonbinary or fluid identities, as well as how identification on such measures would compare to identification on the spectrum measure.

Finally, questions remain on whether different gender identity measures used in the field tap into the same or similar constructs. To illustrate, the current measure used wording around assessing what children *felt* their gender identity to be, whereas other measures have asked simply about what their gender identity is, how similar they are to members of different genders, etc. As one test of the extent to which these different measures are associated with each other, the current work demonstrated relationships of moderate strength, raising the suggestion that the interconnections between these different constructs be further probed. Related to this issue of how gender should be conceptualized and measured by researchers, some have theorized that gender identity is better conceptualized not as a stable trait but as a fluid construct that varies contextually (Mehta, 2015). From this point of view, it is difficult to gauge the stability that should be expected across time in a continuous gender identity measure such as the spectrum. Further investigation of the possibility of contextual effects on gender identification on a spectrum is called for.

A large benefit of the current measure is its ease of use with young samples. Thus, a main area of use for a single-item measure like the spectrum measure could be studies for which gender is not a central measure, as an alternative to asking a categorical question. That said, the current spectrum measure is one of a new generation of measures that aim

to assess gender identity in less categorical ways (e.g., Moore et al., 2020). Thus, it is imperative that researchers continue experimenting with measures that will capture the full spectrum of identities in ways that are responsive both to how children are thinking about their identities and that are developmentally appropriate for the children meant to use them. Additionally, although the current work finds that children's gender identities as measured on the spectrum show moderate stability across a two-year time period, the long-term stability of gender-diverse children's gender identities remains to be studied in depth.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank the following people for their help with data collection, data coding and literature review: Brandon Dull, Rachel Horton, Riley Lowe, Elizabeth Enright, Anne Fast, Daniel Alonso, Jessica Glazier. The authors would also like to thank Dominic Gibson for his assistance with data analyses, and members of the Social Cognitive Development Lab at the University of Washington for their feedback on initial drafts.

Funding.

This research was made available thanks to funding from the National Science Foundation (SMA-1837857 and BCS-1715068) and the National Institute of Child Health and Human Development (HD092347).

References

Bradbard M, & Endsley RC (1983). The effects of sex-Typed labeling on preschool children's information-Seeking and retention. Sex Roles, 9, 247–260. 10.1007/BF00289627

Diamond LM (2020). Gender Fluidity and Nonbinary Gender Identities Among Children and Adolescents. Child Development Perspectives, 14(2), 110–115. 10.1111/cdep.12366

Dinella LM, Fulcher M, & Weisgram ES (2014). Sex-typed personality traits and gender identity as predictors of young adults' career interests. Archives of Sexual Behavior, 43(3), 493–504. 10.1007/s10508-013-0234-6 [PubMed: 24452631]

Egan SK, & Perry DG (2001). Gender identity: a multidimensional analysis with implications for psychosocial adjustment. Developmental Psychology, 37(4), 451–463. doi:10.1037//0012-1649.37.4.451 [PubMed: 11444482]

Etaugh C, Grinnell K, & Etaugh A (1989). Development of gender labeling: Effect of age of pictured children. Sex Roles, 21, 769–773. 10.1007/BF00289807

Fagot BI, Leinbach MD, & O'Boyle C (1992). Gender labeling, gender stereotyping, and parenting behaviors. Developmental Psychology, 28(2), 225–230. 10.1037/0012-1649.28.2.225

Fast AA, & Olson KR (2018). Gender development in transgender preschool children. Child Development, 89(2), 620–637. 10.1111/cdev.12758 [PubMed: 28439873]

Glazier JJ, Gülgöz S, & Olson KR (2020). Gender encoding in gender diverse and gender conforming children. Child Development, 91(6), 1877–1885. 10.1111/cdev.13399 [PubMed: 32686844]

Greenwald AG, Nosek BA, & Banaji MR (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. Journal of Personlaity and Social Psychology, 85(2), 197–216. 10.1037/0022-3514.85.2.197

Gülgöz S, Glazier JJ, Enright EA, Alonso DJ, Durwood LJ, Fast AA, … Olson KR (2019). Similarity in transgender and cisgender children's gender development. Proceedings of the National Academy of Sciences, 116(49), 24480–24485. 10.1073/pnas.1909367116

Halim ML, Ruble D, Tamis-LeMonda C, & Shrout PE (2013). Rigidity in gender-typed behaviors in early childhood: A longitudinal study of ethnic minority children. Child Development, 84(4), 1269–1284. 10.1111/cdev.12057 [PubMed: 23432471]

Hyde JS, Bigler RS, Joel D, Tate CC, & van Anders SM (2019). The future of sex and gender in psychology: Five challenges to the gender binary. American Psychologist, 74(2), 171–193. 10.1037/amp0000307 [PubMed: 30024214]

LaFreniere P, Strayer FF, & Gauthier R (1984). The emergence of same-sex affiliative preferences among preschool peers: A developmental/ethological perspective. Child Development, 55(5), 1958–1965. 10.2307/1129942

Maccoby EE (1998). The two sexes: Growing up apart, coming together. Harvard University Press.

Maccoby EE, & Jacklin CN (1987). Gender segregation in childhood. Advances in Child Development and Behavior, 20, 239–287. 10.1016/S0065-2407(08)60404-8 [PubMed: 3630812]

Martin C, Fabes RA, Evans S, & Wyman H (1999). Social cognition on the playground: Children's beliefs about playing with girls versus boys and their relations to sex segregated play. Journal of Social and Personal Relationships, 16(6), 751–771. 10.1177/0265407599166005

Martin CL, Andrews NC, England DE, Zosuls K, & Ruble DN (2017). A Dual Identity Approach for Conceptualizing and Measuring Children's Gender Identity. Child Development, 88(1), 167–182. 10.1111/cdev.12568 [PubMed: 27246654]

Martin CL, Eisenbud L, & Rose H (1995). Children's gender-based reasoning about toys. Child Development, 66(5), 1453–1471. 10.2307/1131657 [PubMed: 7555224]

Martin CL, & Fabes RA (2001). The stability and consequences of young children's same-sex peer interactions. Developmental Psychology, 37(3), 431–446. 10.1037/0012-1649.37.3.431 [PubMed: 11370917]

McHale SM, Kim JY, Whiteman S, & Crouter AC (2004). Links between sex-typed time use in middle childhood and gender development in early adolescence. Developmental Psychology, 40(5), 868–881. 10.1037/0012-1649.40.5.868 [PubMed: 15355172]

Mehta CM (2015). Gender in context: Considering variability in Wood and Eagly's traditions of gender identity. Sex Roles, 73(11), 490–496. 10.1007/s11199-015-0535-4

Moore JK, Thomas CS, van Hall HW, Strauss P, Saunders LA, Harry M, … & Lin A (2020). The Perth Gender Picture (PGP): Young people's feedback about acceptability and usefulness of a new pictorial and narrative approach to gender identity assessment and exploration. International Journal of Transgender Health, 22(3), 1–12. 10.1080/26895269.2020.1795960

Morgenroth T, & Ryan MK (2020). The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. Perspectives on Psychological Science, 1745691620902442. 10.1177/1745691620902442

Olson KR (2017). When sex and gender collide. Scientific American, 317(3), 44–49. doi:10.1038/scientificamerican0917-44

Olson KR, Key AC, & Eaton NR (2015). Gender cognition in transgender children. Psychological Science, 26(4), 467–474. 10.1177/0956797614568156 [PubMed: 25749700]

Rae JR, Gülgöz S, Durwood L, DeMeules M, Lowe R, Lindquist G, & Olson KR (2019). Predicting Early-Childhood Gender Transitions. Psychological Science, 30(5), 669–681. 10.1177/0956797619830649 [PubMed: 30925121]

Rae JR, & Olson KR (2018). Test-retest reliability and predictive validity of the Implicit Association Test in children. Developmental Psychology, 54(2), 308–330. 10.1037/dev0000437 [PubMed: 29251966]

Reisner SL, Greytak EA, Parsons JT, & Ybarra ML (2015). Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. The Journal of Sex Research, 52(3), 243–256. 10.1080/00224499.2014.886321 [PubMed: 24742006]

Ruble DN, Martin CL, & Berenbaum SA (2006). Gender Development. In Eisenberg N, Damon W, & Lerner RM (Eds.), Handbook of child psychology: Social, emotional, and personality development (pp. 858–932). John Wiley & Sons, Inc.

Salk RH, Thoma BC, & Choukas-Bradley S (2020). The Gender Minority Youth Study: Overview of Methods and Social Media Recruitment of a Nationwide Sample of U.S. Cisgender and Transgender Adolescents. Archives of Sexual Behavior, 49(7), 2601–2610. 10.1007/s10508-020-01695-x [PubMed: 32306108]

Serbin LA, Powlishta KK, & Gulko J (1993). The development of sex typing in middle childhood. Monographs of the Society for Research in Child Development, 58(2), 1–99. 10.2307/1166118

Tobin DD, Menon M, Menon M, Spatta BC, Hodges EV, & Perry DG (2010). The intrapsychics of gender: a model of self-socialization. Psychological Review, 117(2), 601–622. 10.1037/a0018936 [PubMed: 20438239]

Weinraub M, Clemens LP, Sockloff A, Ethridge T, Gracely E, & Myers B (1984). The development of sex role stereotypes in the third year: relationships to gender labeling, gender identity, sex-typed toy preference, and family characteristics. Child Development, 55(4), 1493–1503. 10.2307/1130019 [PubMed: 6488962]

Yunger JL, Carver PR, & Perry DG (2004). Does gender identity influence children's psychological well-being? Developmental Psychoogyl, 40(4), 572–582. 10.1037/0012-1649.40.4.572

Zosuls KM, Ruble DN, Tamis-Lemonda CS, Shrout PE, Bornstein MH, & Greulich FK (2009). The acquisition of gender labels in infancy: implications for gender-typed play. Developmental Psychology, 45(3), 688–701. 10.1037/a0014053 [PubMed: 19413425]

Zosuls KM, Andrews NCZ, Martin CL, England DE, &. Field, R. D. (2016). Developmental changes in the link between gender typicality and peer victimization and exclusion. Sex Roles, 75(5-6), 243–256. 10.1007/s11199-016-0608-z

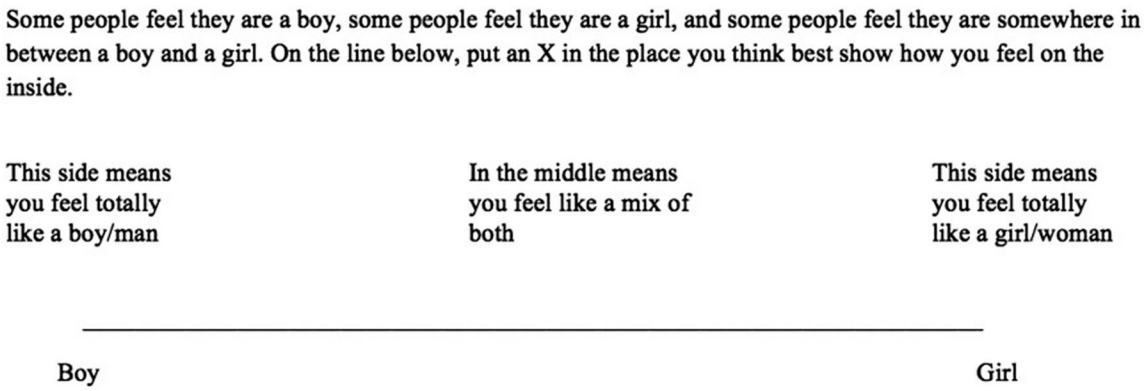

Figure 1. The Gender Spectrum Measure

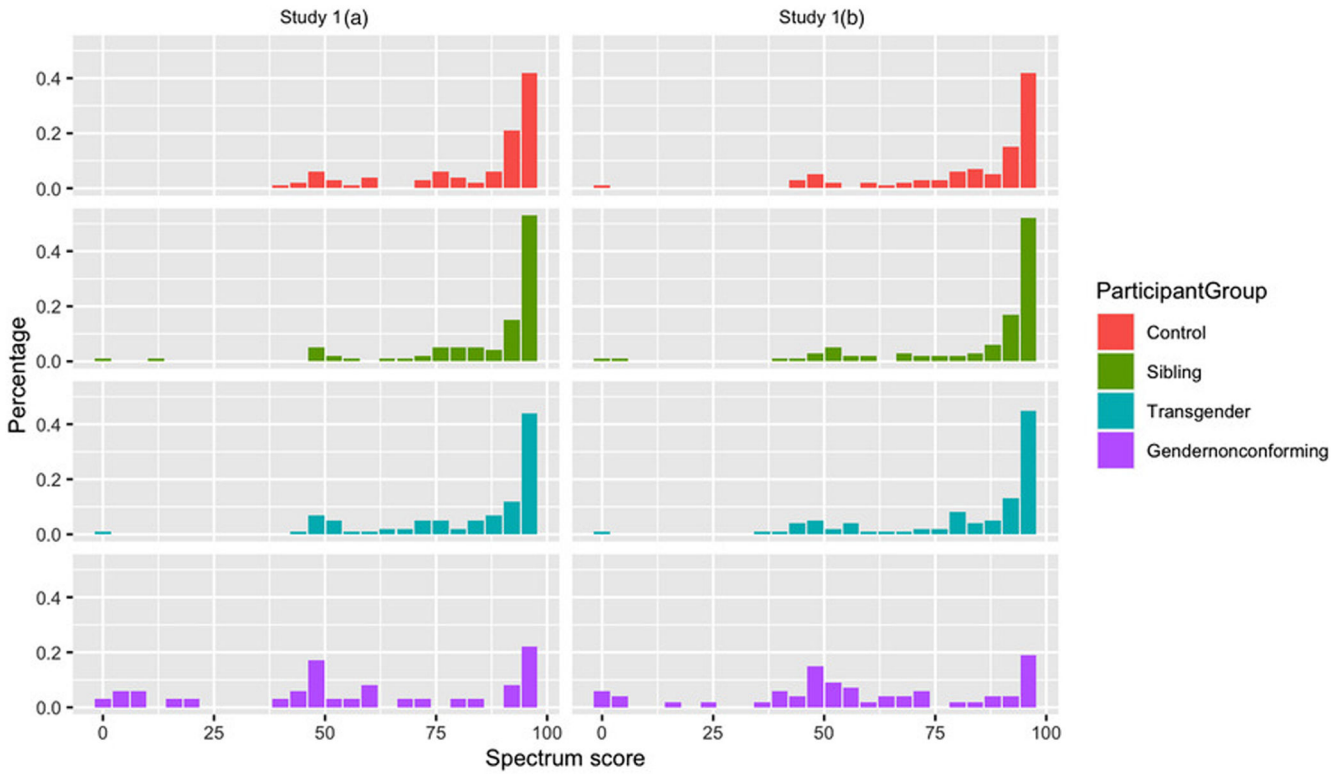

Figure 2. Each group's distribution of spectrum scores in Studies 1a and 1b

Note. For transgender and cisgender participants, spectrum scores are coded so that they range from 0 (feeling totally like the other gender) to 100 (feeling totally like my gender). For gender nonconforming participants, 0 aligns with their sex assigned at birth and 100 aligns with the binary gender that is different from what was assigned at birth.

Table 1

Descriptive statistics for the spectrum measure and comparisons to midpoint and ceiling of scale (Studies 1a and 1b)

	Study 1a			Study 1b		
	***M* (*SD*)**	**Comparison to midpoint**	**Comparison to ceiling**	***M* (*SD*)**	**Comparison to midpoint**	**Comparison to ceiling**
Unrelated cisgender children	86.72 (16.79)	***t*(161) = 27.84, *p* < .001**	***t*(161) = −10.07, *p* < .001**	85.48 (19.49)	***t*(209) = 26.38, *p* < .001**	***t*(209) = −10.80, *p* < .001**
Siblings	89.12 (17.58)	***t*(103) = 22.69, *p* < .001**	***t*(103) = −6.31, *p* < .001**	87.77 (18.85)	***t*(131) = 23.02, *p* < .001**	***t*(131) = −7.45, *p* < .001**
Transgender	85.48 (17.98)	***t*(128) = 22.41, *p* < .001**	***t*(128) = −9.18, *p* < .001**	84.54 (20.08)	***t*(165) = 22.16, *p* < .001**	***t*(165) = −9.93, *p* < .001**
Gender nonconforming	60.71 (31.90)	*t*(35) = 2.01, *p* = .052	***t*(35) = −7.39, *p* < .001**	60.79 (28.61)	***t*(53) = 2.77, *p* = .008**	***t*(53) = −10.07, *p* < .001**

Note. The descriptive statistics shown are the means and standard deviations of each participant group's spectrum scores. For transgender and cisgender participants, spectrum scores ranged from 0 (feeling totally like the other gender) to 100 (feeling totally like my gender). For gender nonconforming participants, 0 aligned with their sex assigned at birth and 100 aligned with a gender that is different from what was assigned at birth. Bolded cells highlight significant *t*-test comparisons.

Table 2

Descriptive statistics for preference, similarity, and implicit gender identity measures (Studies 1a and 1b)

	Study 1a						Study 1b					
	Preferences[1]		Similarity[2]		IAT[3]		Preferences		Similarity		IAT	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Unrelated cisgender children	75.77	14.23	2.03	1.20	0.35	0.44	76.48	15.05	1.94	1.19	0.46	0.47
Siblings	78.12	15.40	1.97	1.52	0.37	0.51	75.22	16.88	2.04	1.27	0.41	0.50
Transgender	76.38	14.48	1.97	1.32	0.27	0.51	72.42	17.26	1.94	1.35	0.34	0.47
Gender nonconforming	69.70	22.52	0.93	1.72	0.16	0.53	60.11	22.93	0.63	1.86	−0.16	0.41
Group comparisons (ANOVA)	$F(3,380) = 2.43$, $p = .065$, $\eta_p^2 = .02$		**$F(3,380) = 6.24$, $p < .001$, $\eta_p^2 = .05$** [4]		$F(3,273) = 1.43$, $p = .234$, $\eta_p^2 = .02$		**$F(3,528) = 12.90$, $p < .001$, $\eta_p^2 = .07$** [5]		**$F(3,538) = 15.70$, $p < .001$, $\eta_p^2 = .08$** [6]		**$F(3,338) = 14.50$, $p < .001$, $\eta_p^2 = .11$** [7]	

Notes.

[1] Preference scores ranged from 0 to 100. For transgender and cisgender participants, 0 indicated preferences aligned with the other gender, 100 indicated preferences aligned with their current gender. For gender nonconforming participants, 0 aligned with their gender assigned at birth, 100 aligned with the other gender.

[2] Similarity scores ranged from 0 to 4. For transgender and cisgender participants, 0 indicated perceived similarity to the other gender, 4 indicated similarity to their current gender. For gender nonconforming participants, 0 aligned with their gender assigned at birth, 4 aligned with the other gender.

[3] For transgender and cisgender participants, negative scores indicated identification with the other gender, and greater positive scores indicated identification with their current gender. For gender nonconforming participants, negative scores showed identification with gender assigned at birth, positive scores showed identification with the other gender.

[4] Post-hoc Tukey comparisons showed that only gender nonconforming children's scores differed from all other groups (ps ≤ .001); all other groups were not significantly different (ps > .980).

[5] Post-hoc Tukey comparisons showed that only gender nonconforming children's scores differed from all other groups (ps < .001); all other groups were not significantly different (ps > .100).

[6] Post-hoc Tukey comparisons showed that only gender nonconforming children's scores differed from all other groups (ps < .001); all other groups were not significantly different (ps > .900).

[7] Post-hoc Tukey comparisons showed that only gender nonconforming children's scores differed from all other groups (ps < .001); all other groups were not significantly different (ps > .200).

Table 3

Correlations between spectrum scores and other measures of gender identity and gender typing

	Study 1a			Study 1b		
	Preferences [1]	**Similarity** [2]	**IAT**	**Preferences** [1]	**Similarity** [2]	**IAT**
Unrelated cisgender children	***r*(142) = .23,*p* = .005**	***r*(143) = .45,*p* < .001**	***r*(104) = −.19, *p* = .046**	***r*(197) = .17,*p* = .017**	***r*(200) = .29,*p* < .001**	*t*(128) = −.02, *p* = .811
Siblings	***r*(88) = .34,*p* = .001**	***r*(87) = .44,*p* < .001**	*t*(67) = −.18, *p* = .150	***r*(117) = .40,*p* < .001**	***r*(120) = .46,*p* < .001**	*t*(79) = .04, *p* = .692
Transgender	*t*(109) = .14, *p* = .138	***r*(109) = .36,*p* < .001**	*t*(79) = −.09, *p* = .412	*t*(152) = .08, *p* = .307	***r*(154) = .19,*p* = .015**	*t*(98) = −.05, *p* = .611

Note. Bolded cells highlight significant correlations with spectrum scores, calculated as identification with own gender.

[1] The preferences scores indicate the extent to which participants showed preferences in toys, clothing, peers associated with their own gender.

[2] The similarity scores were calculated as a difference score of "similarity to own gender" minus "similarity to other gender" on the Martin et al. (2017) gender identity measure.

Table 4

Mean (SD) spectrum scores for each group at each visit

	N	1st spectrum score	2nd spectrum score
Unrelated cisgender children	61	84.90 (18.60)	84.44 (18.17)
Siblings	41	84.24 (22.00)	76.76 (25.91)
Transgender	69	83.62 (20.18)	83.57 (18.63)
Gender-nonconforming	12	44.95 (25.33)	52.95 (18.37)

Note. This table shows descriptive statistics only for participants of Study 2 (i.e., a subset of participants from Studies 1a and 1b who received the spectrum measure twice), and thus reflects data from smaller samples compared to Study 1a and 1b.