Archives of Sexual Behavior

The final publication is available at Springer via http://dx.doi.org/10.1007/s10508-015-0684-0

Original Paper

Factorial Validity and Invariance Assessment of a Short Version of the Recalled Childhood Gender Identity/Role Questionnaire

Jaimie F. Veale

School of Psychology, Te Kura Kete Aronui:Faculty of Arts and Social Sciences, Te Whare Wananga o Waikato: The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand. Email: jveale@waikato.ac.nz

Received: 29 March 2013 / Revised: 31 January 2014 / Accepted: 18 December 2015

DOI: 10.1007/s10508-015-0684-0

ABSTRACT: Recalled childhood gender role/identity is a construct that is related to sexual orientation, abuse, and psychological health. The purpose of this study was to assess the factorial validity of a short version of Zucker et al.'s (2006) "Recalled Childhood Gender Identity/Gender Role Questionnaire" using confirmatory factor analysis and to test the stability of the factor structure across groups (measurement invariance). Six items of the questionnaire were completed online by 1,929 participants from a variety of gender identity and sexual orientation groups. Models of the six items loading onto one factor had poor fit for the data. Items were removed for having a large proportion of error variance. Among birth-assigned females, a five item model had good fit for the data, but there was evidence for differences in scale's factor structure across gender identity, age, level of education, and country groups. Among birth-assigned males, the resulting four-item model did not account for all of the relationship between variables, and modeling for this resulted in a model that was almost saturated. This model also had evidence of measurement variance across gender identity and sexual orientation groups. The models had good reliability and factor score determinacy. These findings suggest that results of previous studies that have assessed recalled childhood gender role/identity may have been susceptible to construct bias due to measurement variance across these groups. Future studies should assess measurement invariance between groups they are comparing and if it is not found the issue can be addressed by removing variant indicators and/or applying a partial invariance model.

KEYWORDS: gender identity; gender role; sexual orientation; confirmatory factor analysis; measurement invariance; psychometric properties

There is a large body of evidence suggesting childhood gender role and identity are related to a number of psychological and sexological outcomes. Due to logistical difficulties of measuring gender role and identity in children and prospectively following them up in adolescence or adulthood, psychometric inventories have been developed to measure adolescents' and adults' recalled childhood gender role and identity. These measures have been used to show that nonconforming recalled childhood gender role/identity is associated with nonheterosexual orientation (e.g. Bailey & Zucker, 1995; Zucker et al., 2006), poorer parental attachment (Landolt, Bartholomew, Saffrey, Oram, & Perlman, 2004), and poorer psychological health (suicidality: Harry, 1983; Plöderl & Fartacek, 2009; anxiety and post-traumatic stress disorder: D'Augelli, Grossman, & Starks, 2006; Lippa, 2008; Roberts, Rosario, Corliss, Koenen, & Austin, 2012; and other psychological health symptoms D'Augelli et al., 2006; Rieger & Savin-Williams, 2012; Skidmore, Linsenmeier, & Bailey, 2006; Weinrich, Atkinson, McCutchan, & Grant, 1995) likely due to this childhood gender nonconformity being associated with abuse and poor treatment in childhood (D'Augelli et al., 2006; Plöderl & Fartacek, 2009).

While measurement of recalled childhood gender role and identity is obviously susceptible to recall bias, evidence from maternal report (Bailey, Miller, & Willerman, 1993), and home videos (Rieger, Linsenmeier, Gygax, & Bailey, 2008) suggests that this bias is not substantial.

Zucker et al.'s (2006) "Recalled Childhood Gender Identity/Gender Role Questionnaire" has been used in a number of studies and is reported to have promising psychometric properties. The original questionnaire had 23 items, with different forms for birth-assigned males and females. Zucker et al. reported two factors emerging from exploratory factor analysis, which they described as Recalled Gender Role/Identity" (18 items) and Closeness to Parents" (3 items). All future studies using the inventory, including this one, have not included the items that loaded on the Closeness to Parents factor. Two further studies have conducted exploratory factor analyses on the questionnaire. On the female version of the questionnaire, Meyer-Bahlburg et al. (2006) reported a solution of three further factors. The largest factor extracted had 13 items—it was labeled Gender Role, and described as approximating the largest factor in the Zucker et al. study. The five items that did not load onto this factor instead loaded onto two other factors labeled Physical Activity (3 items) and Cross-Gender Desire (2 items). Alanko et al. (2008) also assessed the psychometric properties of 13 items of the scale. They suggested a single factor solution based on scree plot observation, but reported two other factors with eigenvalues greater than 1 and three instances of items with standardized factor loadings of .3 or less (two for males and one for females). Plausible explanations for the inconsistent factor analysis findings in these three studies include their exploratory nature and possible differences in the factor structure (invariance) in the different populations sampled in these studies (see Table 1 for details of the samples in these studies).

A number of studies reported internal consistency reliability for the questionnaire. These ranged from .69 to .95 and are outlined in Table 1. One study also reported a "moderate" test-retest correlation of $r = .78$ between test administrations two years apart (Roberts et al., 2012). Given the reported reliability of the scale, a correlation of this magnitude should be interpreted as representing a high amount of stability of reported recalled gender role/identity, with measurement error accounting for most of the variability of scores between time points.

The purpose of this study was to test the validity of a short version of Zucker et al.'s (2006) Recalled Childhood Gender Identity/Gender Role Questionnaire using confirmatory factor analysis. Confirmatory factor analysis has a number of advantages over the exploratory factor analyses used previously: it models measurement error and tests the assumption that this is uncorrelated between items, it allows the testing of hypotheses related to the factor structure, and it provides a stricter test of factorial validity by assessing whether proposed model fit the data.

This study will also assess measurement invariance, by assessing the factorial stability of the Recalled Childhood Gender Identity/Gender Role Questionnaire across a number of demographic groups. The purpose of this is to establish whether the questionnaire is measuring the same construct in different groups. Thus, measurement invariance testing provides an important assessment of the generalizability of reliability and validity findings for the Recalled Childhood Gender Identity/Gender Role Questionnaire across these groups (Vandenberg & Lance, 2000). This testing is conducted within a confirmatory factor analysis framework. It assesses whether the relationship between the questionnaire items and the Recalled Childhood Gender Identity/Role factor is the same across the subgroups, or, in other words, whether there is equivalence of factor structure, item loadings, and intercepts across groups. For example, consider the questionnaire item that asks about preferences for toys and games. If the question itself or the response options (e.g., *very "masculine"*) have different meanings for heterosexuals than for homosexuals, then it is not possible to unambiguously interpret any comparisons between these groups on this item (Vandenberg & Lance, 2000). Any differences (or similarities) found between groups may be due to measurement bias from this different interpretation of the question (Raju, Laffitte, & Byrne, 2002).

There are different levels of measurement invariance that can be tested. 1) *Configural* invariance occurs when the same parameters exist across subgroups (e.g., the same items load onto the same factors in each subgroup). Configural invariance is tested by assessing the overall fit of multiple-group models with the same indicators but parameter estimates (e.g. factor loadings, intercepts) free to differ between groups (Vandenberg & Lance, 2000). 2) *Metric* invariance occurs when factor loadings of items on factors and any item residual correlations do not differ between groups (Vandenberg & Lance, 2000). According to Sass (2011), differences in factor loadings between groups could occur if "the conceptual meaning or understanding…of the construct differs across groups" or "particular items are more applicable for one group than another" (p. 349). 3) A more stringent form of measurement invariance is *scalar* invariance, which occurs when factor loadings, residual correlations, and intercepts of items do not differ between groups (Vandenberg & Lance, 2000). According to Sass (2011), intercept invariance "denotes that subjects with the same latent factor score will have similar responses on average for an item (i.e., observed score) when the latent factor score is zero… [it] could occur due to (a) social desirability reasons or social norms, (b) particular groups displaying a propensity to respond more strongly to an item despite having the same latent trait or factor mean, and/or (c) certain groups having different reference points when making statements about themselves" (p. 349),

Current standards for factorial measurement invariance hold that configural and metric invariance is required for meaningful comparison of factor relationships between groups, and scalar invariance are required for meaningful comparison of mean factor score differences between groups (Chen, 2008; Conroy, Metzler, & Hofer, 2003; Dimitrov, 2010; Gregorich, 2006; Sass, 2011; Steinmetz, 2013).[1] Studies have modified questionnaire items based on measurement invariance findings (e.g., Northrup, Malone, Follingstad, & Stotts, 2013) and this testing has also been applied to groups with large differences in mean scores (e.g.,

Table 1 Details of studies reporting psychometric properties for the Recalled Childhood Gender Role/Identity questionnaire and their internal consistency reliability findings

Study	Sample	Questions used	Cronbach's α
Zucker et al. (2006)	1,305 adolescents and adults, including university students and staff, and a variety of sexual and gender-diverse samples and their family members	18 items extracted from exploratory factor analysis	.92 among all participants
Meyer-Bahlburg et al. (2006)	123 adult women with congenital adrenal hyperplasia and female relatives	13 items from female version extracted from exploratory factor analysis	.90
Alanko et al. (2008)	3,604 Finns recruited from a population twin registry	13 items loading greater than .60 in Zucker et al.'s factor analysis	Males: .69; females: .85
Veale, Clarke, and Lomax (2008)	361 online-recruited male-to-female transsexuals and nontranssexual females	16 items	.90
Plöderl and Fartacek (2009)	Convenience sample of 290 Austrian adults	18 items, translated into German	.95
Roberts et al. (2012)	9,864 young adults from a US population-based cohort	4 items[a]	.74

Note: [a] These items were also used in the present study

Gomez, Vance, & Gomez, 2012; Lavoie & Douglas, 2012; Murray, Booth, McKenzie, Kuenssberg, & O'Donnell, 2014).

Method

Participants and Procedure

Participants were recruited for an internet-based survey described as investigating the development of gender and sexuality. This was conducted through Google advertising to web sites and search pages that had key words such as "transsexual," "transgender," "sexuality," and from contacting international lesbian-, gay-, bisexual-, and/or transgender-related (LGBT) online groups and organizations that had a website asking if they would distribute a call for participants among their members. The call for participants included a brief outline of the aims of the research, what would be involved for participation, and a link for potential participants to access. A press release was also released through Massey University Communications which generated some media attention that is likely to have attracted a significant proportion of the participants with gender-typical identities. Ethical approval for the study was gained through Massey University Human Ethics Committee.

There were 2,709 responses to an online questionnaire. Of these, 196 (7%) could not be used as they did not complete any further than the demographics section at the beginning of the questionnaire. Duplicate responses were identified due to having the same demographic data and occurring within 72 hours. In accordance with the procedure suggested by Bowen, Daniel, Williams, and Baird (2008), 236 responses were deleted because they were duplicates and the more complete response of a duplicate was retained. In all cases, the second version of the duplicate was more complete, suggesting these participants started the questionnaire but had not been able to finish it, and returned later to complete it further. This left a sample of 2,278. Responses were also checked for consistently reporting the same score or extreme scores. None of the responses needed to be removed for not meeting these conditions. Because there were a number of prior questions, 349 participants dropped out of the questionnaire before reaching the Recalled Childhood Gender Identity/Gender Role Questionnaire. This left 1,929 responses on which analyses were conducted.

Table 2 shows participants' gender identity, ethnicity, country, level of education, and age. Participants could select as many of the ethnicity categories as they identified with. Male gender assignment at birth was reported by 1,500 (66%) participants and female gender assignment was reported by 777 (34%). Transsexual participants were those who identified as such. Participants categorized as having an "other gender-variant identity" were those who did not identify as transsexual, but identified with at least one of the other possible gender-variant identities: transvestite, gender queer, drag artist, cross-dresser, androgygne, or bi-, third-, omni-, or non-gendered. Participants who identified as transsexual but didn't report their current gender as opposite to their birth-assigned sex and also identified as a transvestite or cross-dresser were also categorized as having an "other gender-variant identity." The remaining participants categorized as gender-typical did not identify with any of these gender-variant identities. The three levels of gender identity groups were included because there was sufficient sample size to split those with gender-variant identities into two groups, leaving three groups with relatively equal numbers of participants; there was also evidence that the other gender-variant identity group scored distinctly from the other groups on biological factors related to gender identity development (Veale, Clarke, & Lomax, 2010), suggesting differences between transsexual and other gender-variant identity groups should be tested when possible.

Birth-assigned males were overrepresented among participants with gender-variant identities and underrepresented among participants with gender-typical identities. East Asian, Black/African, Māori, and "other ethnic identity" participants were more likely to be birth-assigned female. Participants from the U.S. were more likely to be birth-assigned male and participants from New Zealand were more likely to be birth-assigned female. Birth-assigned male participants were also more likely to report holding a diploma as their highest qualification, and birth-assigned males were significantly older than birth-assigned females, $t(2275) = 20.48$, $p < .001$.

Table 2 Gender identity, education, country, and age of participants, grouped by birth-assigned gender

	Birth-assigned males		Birth-assigned females	
	n	%	*n*	%
Gender identity				
Transsexual	609	41	146	19
Other gender-variant	640	43	259	33
Gender-typical	251	17	372	48
Ethnicity				
White/Caucasian	1,387	93	694	89
East Asian	31	2	36	5
Hispanic/Latino	40	3	26	3
American Indian	39	3	25	3
Black/African	19	1	23	3
South/other Asian	25	2	12	2
Māori	7	1	17	2
Other	14	1	28	4
Country of residence				
USA	906	60	315	41
New Zealand	181	12	246	32
Great Britain	119	8	64	8
Canada	91	6	57	7
Australia	65	4	30	4
Other	138	9	65	8
Level of education				
3 years of high school	97	7	62	8
4 years of high school	168	12	65	9
5 years of high school	147	10	99	13
Diploma	332	23	117	16
Bachelor's degree	423	30	253	34
Master's degree	188	13	123	16
Doctoral degree	79	6	36	5
Age (in years)				
Mean	41.50		29.45	
SD	14.26		11.25	

Note: Other gender-variant identity participants did not identify as transsexual, but identified as at least one of the other possible gender-variant identities.

Participants' sexual orientation was assessed by a single item taken from Bailey (1989), which used a 7-point Kinsey scale to ask if their sexual fantasies were ever about men and/or women. Participants were categorized as gynephilic or androphilic if they reported their sexual fantasies were *always* or *the vast majority* were about women or men, respectively. Participants were categorized as bisexual if their reported sexual fantasies were about women and men equally often, or if they reported "many" sexual fantasies about one gender and "more often" about the other gender. Thirty-three percent of birth-assigned males and 25% of birth-assigned females were categorized as gynephilic, 42% of birth-assigned males and 50% of birth-assigned females were categorized as bisexual, and 25% of birth-assigned males and 25% of birth-assigned females were categorized as androphilic. The 2% of participants who reported not having had sexual fantasies about women or men were excluded from this categorization.

Measure

Zucker et al.'s (2006) Recalled Childhood Gender Identity/Gender Role Questionnaire has 5-point response scales, with one or two further response items to allow respondents to indicate that they did not remember or that the behavior did not apply. All item wordings were preceded with "As a child," and participants were instructed that "Questions that ask about your experiences 'as a child' refer to ages 0 to 12." A shortened version of the questionnaire was administered due to it being part of a large survey (see Veale et al., 2010) and there were concerns about participant fatigue and attrition. Items from Zucker et al.'s questionnaire were selected for inclusion in the study based on having the highest factor loading on Zucker et al.'s recalled childhood gender role/gender identity factor and the overall purposes of the research (see Veale, 2011). The "felt masculinity-femininity," "favorite toys/games," "dress-up play," and "favorite playmates" items were administered on all participants; the "cosmetics/jewelry" and "imitation/admiration of movie/TV characters" items were administered on birth-assigned males only; and the "reputation as a tomboy" and "cross-sex desire private" items were administered on birth-assigned females only. Thus, a total of eight items were assessed, but only six each of these were administered on birth-assigned male and female groups. Some different items were given for each birth-assigned gender because Zucker et al. gave a different version of the questionnaire for males and females. Using different items in this study was permissible because analyses for each birth-assigned gender group were conducted separately and there were no direct comparisons between these groups made in the analysis.

All questions were randomly presented with other items asking about recalled childhood personality, abuse, and anxiety.

Data Analysis

This was conducted using Mplus software version 5.1 (Muthén & Muthén, 2010). Yuan-Bentler robust maximum likelihood method of parameter estimation was used (Satorra & Bentler, 2001; Yuan & Bentler, 2000) as it provides parameter estimates, standard errors, and fit statistics that are robust to missing data and violations of multivariate normality.

Absolute model fit was assessed using the Yuan-Bentler χ^2 likelihood ratio (YBχ^2) and the approximate fit indices: CFI, TLI, RMSEA, and SRMR. An explanation of these fit indices is given in Supplementary Online Materials. A *p* value less than .05 on the χ^2 test indicated model misspecification. On the approximate fit indices, CFI and TLI values greater than .9, RMSEA values less than .05, and SRMR values less than .08 were also used as indicators that the proposed model adequately fitted the data (Kline, 2011).

To detect the parts of the model responsible for misspecification, modification indices were calculated. These are estimates of the expected improvement of model fit from removing a parameter constraint on the model. A modification index score is an approximation of the change on the χ^2 likelihood ratio for the modification at the cost of one degree of freedom (Kline, 2011). Standardized expected parameter change values were also calculated to estimate the magnitude of expected change on the parameter that will result from removing the constraint. In deciding on model modifications, modification indices with the highest values and standardized expected parameter change values of magnitude .20 or greater were used to ensure any change would be of meaningful magnitude. Consideration was also given to whether the modifications made theoretical or conceptual sense (Brown, 2006; Kline, 2011).

Reliability was estimated using Cronbach's (1951) α and Raykov's (1997) factor ρ. The former is an estimate of the intercorrelations between inventory items and the latter is a composite reliability coefficient that is calculated as the ratio of variance explained by the factor to the total variance. Raykov's ρ also has the advantage of incorporating correlated measurement errors in its calculation (Kline, 2011).

Latent factors that are modelled based on observed indicators in factor analysis have the problem of factor score indeterminacy. This means that while the overall structure of the factor can be determined, each case's individual score on the factor cannot be uniquely determined. There are an infinite number of unique scores that each case could have that would be consistent with the factor's structure. The degree of factor indeterminacy is relative, and in situations with higher factor indeterminacy an individual case may be ranked highly relative to others on one set of factor scores and lowly on another, with no way of knowing which of these rankings is "true" (Grice, 2001). Factor score determinacy can be estimated by calculating the squared multiple correlation of the proposed indicators for predicting the Recalled Childhood Gender Identity/Role factor (Mulaik, 2010). This gives us the correlation between the estimated factor scores and the true factor scores (Grice, 2001). It is suggested that this relationship should be of high magnitude. Mulaik (2012) noted that a factor determinacy coefficient of greater than .90 is desirable and less than .71 indicates a severe indeterminacy problem. Reporting factor score determinacy is also useful because while confirmatory factor analysis models with a small number of indicators may be more likely to fit the data, they are also more likely to have factor indeterminacy (Brown, 2006). Thus, both factor score determinacy and model fit can be considered to give counterbalanced indications of the most appropriate number of questionnaire items to retain in a factor model.

The majority of participants sampled in this study lived in two areas: the U.S. (60%) or Australia and New Zealand (20%). Invariance testing was conducted between these groups to test for differences in item meaning for participants living in these regions. The median age of participants was 37 years. Invariance testing was conducted between those above and below the median age, between level of gender identity, sexual orientation (androphilic, bisexual, and gynephilic), and education (3 or 4 years of high school, 5 years of high school or diploma/trade qualification, and university degree).

Invariance testing was conducted by comparing models with invariance constraints to a model without invariance constraints. A statistically significant change in scaled difference χ^2 likelihood ratio test indicates scale measurement variance (Kline, 2011). From findings of simulation studies controlling for sample size, model complexity, and overall model fit, Cheung and Rensvold (2002) suggested a decrease in CFI of less than .01 "indicates that the null hypothesis of invariance should not be rejected" (p. 221). No similar criteria were developed for other fit indices used in this study. Current best practice for assessing whether a scale has measurement invariance is to consider evidence from a range of indicators (Dimitrov, 2010; Kline, 2011; Sass, 2011). Thus, in this study, assessments of invariance were conducted using scaled difference χ^2 likelihood ratio while considering that it can be sensitive to trivial amounts of measurement variance when sample size is large, change in CFI, and an examination of the overall fit for the invariant model.

Participants' responses that they did not engage in this type of play or activity or they do not remember were treated as missing. The percentage of missing data was high (21%), due to participants commonly giving these responses and participant attrition due to the items being spread over a number of pages in the questionnaire (the percentage of missingness varied from 11% to 33%). The missing data handling technique used—Mplus' full information maximum likelihood (Asparouhov & Muthén, 2010)—has showed little bias in estimations and superior performance to other missing data handling techniques in simulation studies when sample size is large (Enders, 2001; Savalei, 2010) even with high proportions of missing data (Scheffer, 2002). The proportion of missing responses to the scale did not differ as a function of sexual orientation, age, gender identity, level of education, or country (see supplementary electronic material).

Results

Because scale items differed between birth-assigned males and females, results are presented separately for these two groups. Item response frequencies and covariance matrices for the six items are given as supplementary electronic material.

Birth-Assigned Males

A model which had all six items loading on a single recalled gender role/identity factor was tested first. As shown in Table 3, this model had good performance on the CFI, TLI, and SRMR fit indices, but evidence of model misspecification was detected on the YBχ^2 likelihood ratio and RMSEA. Examination of the model showed a high percentage of residual error for the "favorite playmates" item (69%) and the "cosmetics" item (52%). A 4-item model with these items removed was tested next. This model had little change in performance on fit indices. The two largest modification indices had standardized expected parameter change values greater than .20. Standardized expected parameter change values for the remaining modification indices were all less than .18. These two modification indices suggested correlated error between the "favorite toys/games" and "imitation/admiration of movie/TV characters" (modification index = 10.2; standardized expected parameter change = .25) and between the "felt masculinity-femininity" and the "dress-up play" items (modification index = 10.2; standardized expected parameter change = .21). Because the former (and not the latter) modification can be conceptualized as due to overlap in item content (playing games), this modification was implemented in the third model presented in Table 3 which had acceptable performance on all fit indices. This 4-item model is illustrated in Fig. 1 and unstandardized parameter estimates are given in supplementary electronic material.

Invariance testing

Table 4 shows the results of invariance testing. The configural invariance model with parameters unconstrained between the three gender identity level groups (transsexual, other gender-variant identity, no gender-variant identity) had no signs of misspecification. When factor loading (metric) invariance was constrained between groups, the overall model fit indices began to show signs of misspecification, there was evidence to reject the null hypothesis of invariance on the scaled-difference χ^2 test, and the change in CFI of .013 just exceeded the .01 criterion suggested by Cheung and Rensvold (2002).

When intercepts were also constrained (scalar invariance) between gender identity groups, there was strong evidence to reject the invariance hypothesis: the scaled-difference χ^2 test was highly statistically significant, the change in CFI was well above the .01 criterion, and the overall model fit became notably worse on all indicators.

For sexual orientation, configural invariance was established and there was no notable worsening of model fit detected on the

Table 3 Fit statistics, reliability, and factor score determinacy estimates for models of the Recalled Childhood Gender Identity/Role Inventory in birth-assigned males

Model	YBχ^2	*d*	*p*	CFI	TLI	RMSEA	SRM	ρ	α	FSD
6 items	50.55	9	< .001	.98	.96	.060	.03	.86	.88	.94
4 items	11.19	2	.004	.99	.98	.060	.02	.85	.87	.93
4 items, Favorite toys ⌣ Imitation TV characters	0.78	1	.376	1.00	1.00	.000	.00	.83	.87	.92

Note: N = 1,263; YBχ^2 = Yuan-Bentler χ^2 likelihood ratio; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; see supplementary electronic material for explanation of fit indices; ρ = Raykov's composite reliability; FSD = factor score determinacy; ⌣ error correlation between items

Table 4 Invariance testing fit statistics for the final Recalled Childhood Gender Identity/Role model in birth-assigned males

Model	$YB\chi^2$	*df*	*p*	$YB\chi^2_{SD}$	Δ*df*	*p*	CFI	ΔCFI	RMSEA
Gender identity, *n* = 1263									
Configural invariance	2.38	3	.497	-	-	-	1.000	-	.000
Metric invariance	19.53	9	.021	4.61	6	.026	.987	.013	.053
Scalar invariance	60.16	15	< .001	48.35	12	< .001	.944	.056	.085
Sexual orientation, *n* = 1215									
Configural invariance	3.63	3	.304	-	-	-	.999	-	.023
Metric invariance	11.13	9	.267	6.69	6	.350	.998	.001	.024
Scalar invariance	64.83	15	< .001	54.64	12	< .001	.953	.046	.091
Country: USA/Australia or New Zealand, *n =1058*									
Configural invariance	2.50	2	.287	-	-	-	.999	-	.022
Metric invariance	10.70	5	.058	5.39	3	.145	.994	.005	.046
Scalar invariance	11.43	8	.178	5.87	6	.118	.996	.003	.028
Age (median split), *n = 1247*									
Configural invariance	1.55	2	.388	-	-	-	1.000	-	.000
Metric invariance	3.76	5	.584	1.87	3	.600	1.000	.000	.000
Scalar invariance	6.19	8	.627	3.92	6	.688	1.000	.000	.000
Level of education, *n = 1206*									
Configural invariance	1.07	3	.786	-	-	-	1.000	-	.000
Metric invariance	3.08	9	.961	1.70	6	.945	1.000	.000	.000
Scalar invariance	11.88	15	.688	9.13	12	.691	1.000	.000	.000

Note. $YB\chi^2$ = Yuan-Bentler χ^2 likelihood ratio; SD = scaled difference; *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation

metric invariance model. In the scalar invariance model, there was strong evidence to reject the invariance hypothesis: the scaled-difference χ^2 test was highly statistically significant, the change in CFI was well above the .01 criterion, and the overall model fit became notably worse on all indicators.

For country, age, and level of education, there configural invariance was established across all subgroups and there was no notable worsening of model fit detected on any of the metric or scalar invariance models.

Table 5 outlines modification indices for models with misspecification detected: models of degree of gender identity and sexual orientation groups with factor loading and intercept (scalar) invariance constrained between. Cronbach's alpha reliability coefficients for each of the subgroups are given in supplementary electronic material.

Birth-Assigned Females

As outlined in Table 6, a model with all six items loading on a factor had evidence of model misspecification on the $YB\chi^2$ likelihood ratio but good performance on the other fit indices. Examination of the model showed percentage of residual error for the "favorite playmates" item was 62%. All other items were modeled with residual error of 39% or less. A 5-item model with the "favorite playmates" item removed was tested and this did not show evidence of misspecification on any of the fit indices. This 5-item model is illustrated in Figure 2 and unstandardized parameter estimates are given in supplementary electronic material.

Invariance testing

Table 7 shows the results of invariance testing for birth-assigned females. There appeared to be configural invariance between gender identity groups—while the RMSEA statistic was marginally greater than the .05 criterion, the $YB\chi^2$ test and CFI were within acceptable range to not reject the null hypothesis of configural invariance. When metric invariance was constrained between gender identity groups, there was worsening of model fit—both the scaled-difference χ^2 test and change in CFI just reached the criteria needed to reject the null hypothesis of metric invariance and the overall model began to show signs of misspecification on the $YB\chi^2$ and RMSEA indices. When scalar invariance was constrained between gender identity groups, the model fit became progressively worse on all indicators.

Table 5 Modification indices for the scalar invariance degree of gender identity and sexual orientation group models for birth-assigned males.

	Group to be modified	Modification index	Standardized expected parameter change
Degree of gender identity scalar variance			
"Felt masculinity-femininity" intercept	NGV	17.6	.16
"Dress-up play" intercept	OGV	17.4	.14
	NGV	15.6	-.19
"Admiration movie/TV chars." intercept	Transsexuals	11.7	.24
Sexual orientation scalar invariance			
"Favorite toys" intercept	Androphilic	42.3	.27
	Gynephilic	14.0	-.08
"Dress-up play" intercept	Androphilic	10.0	-.14

Note: OGV = other gender-variant identity; NGV = no gender-variant identity

Table 6 Fit statistics, reliability, and factor score determinacy estimates for models of the Recalled Childhood Gender Identity/Role Inventory in birth-assigned females

Model	YBχ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR	ρ	α	FSD
6 items	21.75	9	.001	.99	.99	.046	.02	.91	.92	.96
5 items	8.04	5	.154	1.00	1.00	.030	.01	.92	.94	.96

Note. N = 666; YBχ^2 = Yuan-Bentler χ^2 likelihood ratio; df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; ρ = Raykov's composite reliability; FSD = factor score determinacy

Across sexual orientation groups, configural and metric invariance was established with no signs of model misspecification. The change in CFI for the scalar invariance model of .009 was just under the .01 criterion, the scaled difference χ^2 test was not statistically significant though, and the overall model fit was within the acceptable bounds of all indicators so the null hypothesis of scalar invariance was not rejected.

For country, configural and metric invariance models did not show indications of misspecification, but a case could be made for rejecting the scalar invariance model based on a statistically significant scaled-difference χ^2 test and a change in CFI of .009, just under the .01 criterion. The overall country scalar invariance model also showed some signs of misspecification on the YBχ^2 and RMSEA indices.

For models with age (median split) and level of education groups, the configural (unconstrained) models had signs of misspecification on the YBχ^2 likelihood ratio and RMSEA, indicating different model parameters across these groups. There was no notable worsening of model fit detected on either the metric or scalar invariance models for age (median split). For level of education groups, constraining metric invariance did not notably alter model fit, but there was some evidence to reject the scalar invariance hypothesis: the scaled-difference χ^2 test was statistically significant, the change in CFI was greater than the .01 criterion, and the overall fit of the model became somewhat worse. Table 8 shows modification indices for these models that had misspecification. While a decision on whether the scalar invariance between country groups would be marginal, modification indices for this model are shown anyway. Cronbach's alpha reliability coefficients for each of the subgroups are given in supplementary electronic material.

Discussion

The aims of this study were to assess the factorial validity of a short version of the Recalled Childhood Gender Identity/Gender Role Questionnaire using confirmatory factor analysis and to test the measurement stability of the scale across a number of key demographic groups. This assessment was conducted on a large online community sample. Participation was open to any adult, and the recruitment strategy and project topic meant there was a large proportion of lesbian, gay, bisexual, and transgender participants. The diversity of reported sexual orientation in this sample was also consistent with other studies of transgender people (e.g., Nieder et al., 2011; Nuttbrock et al., 2011).

Zucker et al. (2006) included a total of 18 items with standardized factor loadings greater than .40 in a single factor from an exploratory factor analysis for the questionnaire. This study assessed eight of these items, two on birth-assigned males only, two on birth-assigned females only, and four on all participants. Of these eight items, two had high proportions of error variance. Less than half of the variance of the "favorite playmates" and "cosmetics" items could be attributed to the recalled childhood gender identity/role latent factor in models with birth-assigned males and females.[2] The factor loadings found in this study were of similar magnitude to those found in previous exploratory factor analyses (Meyer-Bahlburg et al., 2006; Zucker et al., 2006). While it is common to include items with standardized factor loadings of

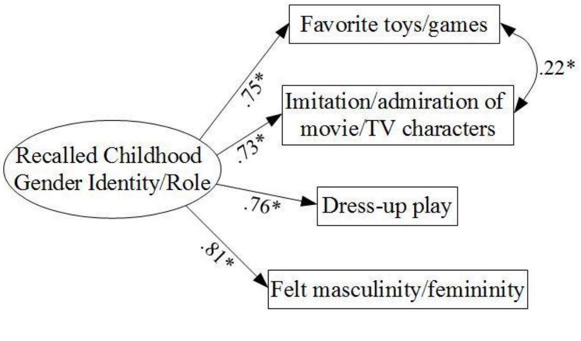
Favorite toys/games
Imitation/admiration of movie/TV characters
.22*
.75*
.73*
Recalled Childhood Gender Identity/Role
.76*
Dress-up play
.81*
Felt masculinity/femininity

Figure 1 The final 4-item factor model for birth-assigned males with standardized parameter estimates. * *p* < .001.

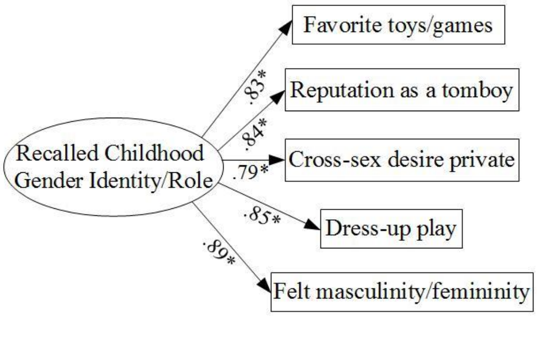

Figure 2 The final 5-item factor model for birth-assigned females with standardized parameter estimates. * *p* < .001.

Table 7 Invariance testing fit statistics for the final Recalled Childhood Gender Identity/Role model in birth-assigned females

Model	YBχ^2	*df*	*p*	YBχ^2_{SD}	Δ*df*	*p*	CFI	ΔCFI	RMSEA
Gender identity, *n* = 666									
Configural invariance	23.63	15	.072	-	-	-	.991	-	.051
Metric invariance	40.17	23	.017	15.50	8	.051	.981	.010	.058
Scalar invariance	118.38	31	< .001	88.81	16	< .001	.906	.085	.113
Sexual orientation, *n* = 640									
Configural invariance	13.61	15	.555	-	-	-	1.000	-	.000
Metric invariance	29.79	23	.155	14.79	8	.063	.994	.006	.037
Scalar invariance	41.57	31	.097	25.56	16	.061	.991	.009	.040
Country: USA/Australia or New Zealand, *n* =565									
Configural invariance	14.94	10	.134	-	-	-	.996	-	.042
Metric invariance	20.59	14	.113	5.24	4	.264	.994	.002	.041
Scalar invariance	32.62	18	.019	16.69	8	.037	.987	.009	.054
Age (median split), *n* = 562									
Configural invariance	21.78	10	.016	-	-	-	.991	-	.060
Metric invariance	25.08	14	.034	3.05	4	.549	.992	-.001	.049
Scalar invariance	28.43	18	.056	6.14	8	.631	.992	-.001	.042
Level of education, *n* = 647									
Configural invariance	27.60	15	.024	-	-	-	.990	-	.062
Metric invariance	44.01	23	.005	14.49	8	.070	.984	.006	.065
Scalar invariance	63.04	31	.001	31.28	16	.012	.976	.014	.069

Note. YBχ^2 = Yuan-Bentler χ^2 likelihood ratio; SD = scaled difference; *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation

.40 or greater in exploratory factor analysis as Zucker et al. did, in confirmatory factor analysis it is common to remove items with loadings less than .70 (Kline, 2011). This is because items need to have a standardized factor loading of .7 or greater to have 50% of their variance accounted for by the factor. While allowing items with a larger proportion their variance measuring something other than the Recalled Childhood Gender Role/Identity factor (measurement error) may be considered justified due to the recalled nature of the construct, this and other studies have found a range of items with high factor loadings, suggesting this doesn't need to be the case. Allowing items with factor loadings lower than .7 in confirmatory factor analysis and structural equation models can also make these models susceptible to inadmissible estimates of negative variance or correlations greater than 1 (Kline, 2011). Lower factor loadings are also related to lower reliability and lower factor score determinacy, which results in imprecision in individual scores on this construct. The finding of lower factor loadings for some items does not necessarily mean they are poor at measuring recalled childhood gender role/identity. It could also be due to existence of an alternative factor structure—a second-order factor model or a model with separate but related factors (cf. Meyer-Bahlburg et al., 2006). Future research with a longer version of the questionnaire could assess the relative fit of models with these different factor structures.

While shorter versions of a questionnaires are more convenient to administer and less susceptible to fatigue effects, it is important to ensure adequate reliability and coverage of the construct are retained. The final models in this study had good reliability (α = .87 and .94) and this was comparable to studies that have used longer versions of the questionnaire (see Table 1). The factors also had high factor score determinacy indicators, which is important because factor indeterminacy can be a concern in factors with a small number of indicators for factor estimation (Williams, 1978). The questions retained in the final models also covered a varied range of topics related to gender role and identity, suggesting coverage of the breadth of the construct. This finding provides preliminary evidence that researchers interested in using a brief version of the Recalled Childhood Gender Role/Identity Questionnaire could be justified in using a smaller set of items such as the items that were assessed in this study. The results of this study were, however, limited by the fact that data were not collected for the entire questionnaire to allow comparisons between the performance of the shorter and the full length versions of the questionnaire.

Among birth-assigned males, modifications resulted in a 4-item model that still did not appear to fit the data adequately, likely due to not accounting for all of the correlation between the "favorite toys/games" and "imitation/admiration of movie/TV characters" items. It is possible that there is some aspect of the relationship between these two items, related to game/play preference, that is not accounted for by the Childhood Gender Role/Identity factor. This may also be indicative of a second order factor existing in the measurement model. When error variance between these two items was allowed to correlate, the model fit indices became acceptable, but the resulting model had only one remaining degree of freedom, making it limited in the number of elements in which it could possibly be rejected and less likely to be replicable across samples (Kline, 2011).

Invariance testing among birth-assigned males suggested the same single-factor configuration held across subgroups. Imposing parameter constraints elicited evidence of factorial instability across both gender identity and sexual orientation groups. Modification indices for these models suggested that intercepts for a number of the items differed among these subgroups. As outlined in Table 5, almost all of the items and gender identity and sexual orientation subgroups were represented in the modification indices, and there was no discernable pattern to these. While these findings may have been caused by both sexual orientation and gender identity measurement variance, it is also possible that the gender identity measurement variance findings were due to the

Table 8 Modification indices for multi-group models with misspecification for birth-assigned females.

	Group to be modified	Modification index	SEPC
Degree of gender identity scalar invariance			
“Cross-sex desire private” factor loading	Transsexual	36.8	.16
	NGV	45.2	-28
“Cross-sex desire private” intercept	Transsexual	46.4	-.65
	NGV	35.1	.09
“Reputation as a tomboy” factor loading	Transsexual	13.0	.17
	NGV	19.7	-.10
“Reputation as a tomboy” intercept	Transsexual	16.6	.39
“Favorite toys” intercept	NGV	12.5	-.07
“Reputation as a tomboy” ↔ “felt masculinity-femininity”	NGV	9.8	.32
“Reputation as a tomboy” ↔ “cross-sex desire private”	NGV	7.8	-.25
“Reputation as a tomboy” ↔ “dress-up play”	Transsexual	7.7	.44
Country scalar invariance			
“Favorite toys” intercept	Australia/NZ	6.9	-.13
“Cross-sex desire private” intercept	Australia/NZ	6.5	.11
Age (median split) configural invariance			
“Cross-sex desire private” ↔ “dress-up play”	Younger	9.6	.25
“Reputation as a tomboy” ↔ “dress-up play”	Older	6.5	.38
Level of education configural invariance			
“Cross-sex desire private” ↔ “favorite toys”	5 years high school/diploma/trade qual.	8.5	-.33
“Reputation as a tomboy” ↔ “cross-sex desire private”	University degree	7.7	-.25
Level of education scalar invariance			
“Reputation as a tomboy” factor loading	University degree	13.2	.09
	5 years high school/diploma/trade qual.	11.6	-.16
“Cross-sex desire private” intercept	3-4 years high school	9.1	-.18
“Reputation as a tomboy” ↔ “cross-sex desire private”	University degree	7.8	-.23
“Cross-sex desire private” ↔ “favorite toys”	5 years high school/diploma/trade qual.	6.0	-.25

Note. SEPC = standardized expected parameter change; NGV = no gender-variant identity; ↔ error correlation between items

different distribution of the sexual orientation subgroups across the gender identity subgroups. Likewise, it is possible that the sexual orientation measurement variance findings were due to the different distribution of the levels of gender identity across the sexual orientation subgroups. The current study lacked the sample size to be able to split the birth-assigned males into nine subgroups to be able to clarify this, making it an issue for future research—perhaps most easily addressed by replicating the current analyses on nontransgender participants. There was no evidence of factorial instability across country, age, and level of education subgroups among birth-assigned males, so the instability across gender identity and sexual orientation subgroups could not be attributed to differences in the distribution of these other demographics across the gender identity and sexual orientation subgroups. The sample was also overwhelmingly white (92%), so it is also unlikely that ethnicity differences in these groups could explain this measurement variance either.

For birth-assigned females, the 5-item model showed acceptable model fit when all participants were tested as one group and had good performance on reliability and factor score determinacy indicators. Reliability estimates were higher for birth-assigned females than their opposite assigned-gender counterparts, likely due to the final scale having an additional item in this group. The model did not fare so well when birth-assigned females were split into subgroups. Configural invariance models of age groups (median split) and level of education groups had signs of model misspecification, suggesting different factor configurations in these subgroups. Modification indices suggested the Childhood Gender Role/Identity factor did not accurately account for the correlation between items among different subgroups (see Table 8 for details). Constraining factorial invariance between gender identity, country, and level of education subgroups also resulted in significant worsening of model fit. Modification indices for the invariant gender identity and level of education models suggested removing constraints on various factor loadings and intercepts and allowing error covariance between many of the items across most of the groups. For the country invariance model, modification indices were limited to intercepts in the Australia/New Zealand group. As with birth-assigned males, the precise groups that are not measurement invariant is uncertain because it could be accounted for by the unequal composition of some other demographic group which is not invariant in each of the subgroups which could confound the group comparisons. It can be concluded that there is measurement variance in at least some of these subgroups though.

In sum, these findings suggest that there is substantial evidence to reject the hypothesis that the Recalled Childhood Gender Identity/Gender Role Questionnaire measures a single latent factor in a way that is stable across groups. For birth-assigned males, there were problems with the model not accounting for all of the relationship between variables, and modeling the resultant correlated error variance left a model that was almost saturated, using up almost all of the model’s degrees of freedom. There was also evidence for some item intercept instability (noninvariance) across gender identity and sexual orientation groups. For birth-assigned females, there was evidence for some item intercept, factor loading, and error covariance instability across a wide range of groups.

The evidence of factorial measurement instability across these groups creates concern for the validity of the questionnaire in samples that span or compare gender identities, sexual orientations, age groups, and levels of education. Specifically, configural and metric invariance are required for meaningful comparison of factor relationships between groups and item intercept (scalar) invariance is further required for meaningful comparison of mean score differences between groups (Conroy et al., 2003; Gregorich, 2006). A number of studies have used the Recalled Childhood Gender Identity/Gender Role Questionnaire on samples that have spanned the groups that were not found to be invariant in this study. Some studies have also directly assessed this inventory's relationship with sexual orientation (Rieger & Savin-Williams, 2012), whether gender moderates the inventory's relationships with other variables (Alanko et al., 2008; Roberts, Rosario, Slopen, Calzo, & Austin, 2013), and compared means of gender identity and sexual orientation groups (Drummond, Bradley, Peterson-Badali, & Zucker, 2008; Plöderl & Fartacek, 2009; Singh et al., 2010; Singh, McMain, & Zucker, 2011; Veale, Clarke, & Lomax, 2008; Zucker et al., 2006, 2012). A large number of studies have consistently found sexual orientation differences on recalled childhood gender role and identity using Zucker et al.'s (2006) questionnaire and other questionnaires, with homosexuals having a more gender-atypical mean score (see Bailey & Zucker, 1995; and Zucker et al., 2006 for reviews). These studies have assumed that the instruments used assess an equivalent construct across groups. The findings of this study suggest that this assumption should be tested.

Studies of recalled childhood gender role/identity may be subject to construct bias. This study's findings of measurement variance suggest the meaning of the underlying recalled childhood gender role/identity latent construct may differ across these groups. This difference in meaning could, at least partially, explain any group-difference findings. The findings of this study may also generalize to other studies that have used different questionnaires to measure recalled childhood gender role/identity as the items used in this study are similar to those used in other questionnaires measuring this construct. This finding underscores the importance of using multiple methods to measure group differences in this construct, including prospective studies (Steensma, Biemond, de Boer, & Cohen-Kettenis, 2011) and home videos (Rieger et al., 2008).[3]

Interpretation of the extent of the measurement variance in these models involved examining the overall fit of the models and the change in fit resulting from invariance constraints. A statistically significant p value on the Yuan-Bentler χ^2 test corresponded to a rejection of the null hypothesis of exact model fit or the same model fit on the scaled-difference χ^2 test. A change in CFI values corresponds to the amount of change of model fit on a scale from an exactly fitting model (CFI of 1) to the least well-fitting model with all parameters constrained to 0 (CFI of 0). A RMSEA score of .05 or greater can be interpreted as a rejection of the "close" fit hypothesis (Kline, 2011; see supplementary materials for more details of these fit indices). Standardized expected parameter change scores given in Tables 5 and 8 also give estimations of the likely direction and magnitude of parameter differences across groups. These parameters were small to moderate in magnitude and they did not manifest in a consistent direction, such as one group seeming to consistently have higher factor loadings or intercepts. If that had have been the case, then it would have been possible to draw upon other work to refine this study's conclusion to suggest that differences between any groups would be likely to be overestimated or underestimated. From simulation studies, Chen (2008) found that when parameters that are not invariant are not in a consistent direction then the construct bias is not as pronounced as if these are in a consistent direction. Chen noted, however, that the construct validity was still questionable because different constructs are still being measured across different groups.

Future studies of group differences in recalled childhood gender role/identity should test for measurement invariance to ensure any group differences cannot be explained by the construct being measured differently between these groups. If the findings of this study are replicated, confirming this measurement inconsistency across groups (noninvariance) then researchers are faced with a decision on how to proceed. One option is to utilize a partial invariance model in which some parameters that are not invariant are allowed to differ, thus accounting for the measurement variance within the model. This results in factors that are not exactly comparable across groups, but for practical purposes allowing some parameter differences across groups may only have negligible impact on factor comparability (Sass, 2011). There are, however, no agreed-upon guidelines for the proportion of parameters that must be invariant for groups to be able to be meaningfully compared in partial invariance models (Vandenberg & Lance, 2000). Some have suggested this be in excess of 80% (Byrne, Shavelson, & Muthén, 1989; Dimitrov, 2010) or 50% as long as there is an adequate theoretical basis for allowing the parameters to be unconstrained (Vandenberg & Lance, 2000). Regardless, recent simulation research has shown that summed scale scores should not be used in situations of partial invariance of even one variant indicator intercept (Steinmetz, 2013). Another option is to remove items that are not invariant from the model. In considering this, researchers should assess whether they retain adequate coverage of the construct within the remaining items, and be aware that creating different versions of the questionnaire to assess differences is less practical and would reduce comparability across studies. These decisions should be based on the types of analyses the researcher is attempting, the amount of invariance they have in their study, and the number of items they used to measure recalled childhood gender role/identity. Until a short-form that is robust to group measurement differences is created, it is suggested that researchers use a longer version of the measure, especially if they want to have the option of excluding items that are not measurement invariant from their calculation of the construct.

The findings of this study uncover a number of avenues for future research. Using more items from the questionnaire would also hopefully negate problem of an almost saturated model that was encountered in this study. This study was conducted on a convenience sample with different recruitment strategies targeting LGBT and heterosexual participants. A more representative sample or a sample with heterosexual and LGBT participants matched would allow more robust results[2]. It has also been suggested that 5-point response Likert scales could be treated as ordinal, rather than continuous in confirmatory factor analysis (Kline, 2011). Modelling this inventory as ordinal with polychoric correlations would also allow examination of item response thresholds. Because the questions used in this study were selected ad hoc based on the needs of the study, the results of this study should not be the basis for selecting a short version of the Recalled Childhood Gender Identity/Gender Role Questionnaire. It would be useful to develop such a questionnaire, however, and the findings of this and previous factor analysis studies could potentially assist with this. Moreover, this study indicated that measurement invariance across groups should be assessed when considering which questions to retain.

Conclusions

Using confirmatory factor analysis and invariance testing, this study found areas of concern for a short version of the Recalled Childhood Gender Role/Identity Questionnaire. Although it has been established that the questionnaire can predict gender identity and sexual orientation with some accuracy, the question assessed in the current study is whether this can be attributed to an underlying Recalled Childhood Gender Role/Identity factor or to differences in response biases, such as one group tending to agree more to the questions, or different subgroups interpreting the items differently because the subgroups' divergent experiences making them tend to give different meaning to the questions (Steinmetz, 2013). It seems unlikely that all of the differences between groups on the Recalled Childhood Gender Role/Identity questionnaire can be attributed to differences in biases and not differences in the underlying construct. However, this study found that this measure seems to be susceptible to some measurement bias differences across groups and researchers who are interested in measuring the magnitude of group differences on this construct should be aware of this issue and use items that have measurement invariance across the groups that they are comparing or account for any biases using a partial invariance model.

References

Alanko, K., Santtila, P., Harlaar, N., Witting, K., Varjonen, M., Jern, P., … Sandnabba, N. K. (2008). The association between childhood gender atypical behavior and adult psychiatric symptoms is moderated by parenting style. *Sex Roles*, *58*, 837–847. http://doi.org/10.1007/s11199-008-9395-5

Asparouhov, T., & Muthén, B. O. (2010, September 29). *Multiple imputation with Mplus.* Retrieved from http://statmodel.com/download/Imputations7.pdf

Bailey, J. M. (1989). A test of the maternal stress hypothesis for human male homosexuality. *Dissertation Abstracts International, 50*(10), 4761B.

Bailey, J. M., Miller, J. S., & Willerman, L. (1993). Maternally rated childhood gender nonconformity in homosexuals and heterosexuals. *Archives of Sexual Behavior*, *22*, 461–469.

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

Bowen, A. M., Daniel, C. M., Williams, M. L., & Baird, G. L. (2008). Identifying multiple submissions in internet research: Preserving data integrity. *AIDS and Behavior*, *12*, 964–973. http://doi.org/10.1007/s10461-007-9352-2

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466. http://doi.org/10.1037/0033-2909.105.3.456

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005–1018. http://doi.org/10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Conroy, D. E., Metzler, J. N., & Hofer, S. M. (2003). Factorial invariance and latent mean stability of performance failure appraisals. *Structural Equation Modeling*, *10*, 401–422. http://doi.org/10.1207/S15328007SEM1003_4

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. http://doi.org/10.1007/BF02310555

D'Augelli, A. R., Grossman, A. H., & Starks, M. T. (2006). Childhood gender atypicality, victimization, and PTSD among lesbian, gay, and bisexual youth. *Journal of Interpersonal Violence*, *21*, 1462–1482. http://doi.org/10.1177/0886260506293482

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121–149. http://doi.org/10.1177/0748175610373459

Drummond, K. D., Bradley, S. J., Peterson-Badali, M., & Zucker, K. J. (2008). A follow-up study of girls with gender identity disorder. *Developmental Psychology*, *44*, 34–45.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370. http://doi.org/10.1037/1082-989X.6.4.352

Gomez, R., Vance, A., & Gomez, A. (2012). Children's Depression Inventory: Invariance across children and adolescents with and without depressive disorders. *Psychological Assessment, 24*, 1–10. http://doi.org/10.1037/a0024966

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), S78–S94. doi:10.1097/01.mlr.0000245454.12228.8f

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450. http://doi.org/10.1037/1082-989X.6.4.430

Harry, J. (1983). Parasuicide, gender, and gender deviance. *Journal of Health and Social Behavior*, *24*, 350–361.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Landolt, M. A., Bartholomew, K., Saffrey, C., Oram, D., & Perlman, D. (2004). Gender nonconformity, childhood rejection, and adult attachment: A study of gay men. *Archives of Sexual Behavior*, *33*, 117–128. http://doi.org/10.1023/B:ASEB.0000014326.64934.50

Lavoie, J. A. A., & Douglas, K. S. (2012). The perceived stress scale: Evaluating configural, metric and scalar invariance across mental health status and gender. *Journal of Psychopathology and Behavioral Assessment*, *34*, 48–57. http://doi.org/10.1007/s10862-011-9266-1

Lippa, R. A. (2008). The relation between childhood gender nonconformity and adult masculinity–femininity and anxiety in heterosexual and homosexual men and women. *Sex Roles*, *59*, 684–693. http://doi.org/10.1007/s11199-008-9476-5

Meyer-Bahlburg, H. F. L., Dolezal, C., Zucker, K. J., Kessler, S. J., Schober, J. M., & New, M. I. (2006). The Recalled Childhood Gender Questionnaire-Revised: A psychometric analysis in a sample of women with congenital adrenal hyperplasia. *Journal of Sex Research*, *43*, 364–367. http://doi.org/10.1080/00224490609552335

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Mulaik, S. A. (2012, September 30). *Re: How to report SEM results: Power & effect size* [Electronic mailing list message]. Retrieved from https://listserv.ua.edu/cgi-bin/wa?A2=ind1209&L=semnet&F=&S=&P=283734

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are autistic traits measured equivalently in individuals with and without an autism spectrum disorder? An invariance analysis of the Autism Spectrum Quotient Short Form. *Journal of Autism and Developmental Disorders*, *44*, 55–64. http://doi.org/10.1007/s10803-013-1851-6

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Nieder, T. O., Herff, M., Cerwenka, S., Preuss, W. F., Cohen-Kettenis, P. T., De Cuypere, G., … Richter-Appelt, H. (2011). Age of onset and sexual orientation in transsexual males and females. *Journal of Sexual Medicine*, *8*, 783–791. http://doi.org/10.1111/j.1743-6109.2010.02142.x

Northrup, T. F., Malone, P. S., Follingstad, D., & Stotts, A. L. (2013). Using item response theory to improve alcohol dependence screening for African American and White male and female college students. *Addictive Disorders and Their Treatment*, *12*, 99–109. http://doi.org/10.1097/ADT.0b013e3182627431

Nuttbrock, L., Bockting, W., Mason, M., Hwahng, S., Rosenblum, A., Macri, M., & Becker, J. (2011). A further assessment of Blanchard's typology of homosexual versus non-homosexual or autogynephilic gender dysphoria. *Archives of Sexual Behavior*, *40*, 247–257. http://doi.org/10.1007/s10508-009-9579-2

Plöderl, M., & Fartacek, R. (2009). Childhood gender nonconformity and harassment as predictors of suicidality among gay, lesbian, bisexual, and heterosexual austrians. *Archives of Sexual Behavior*, *38*, 400–410. http://doi.org/10.1007/s10508-007-9244-6

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529. http://doi.org/10.1037//0021-9010.87.3.517

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173–184. http://doi.org/10.1177/01466216970212006

Rieger, G., Linsenmeier, J. A. W., Gygax, L., & Bailey, J. M. (2008). Sexual orientation and childhood gender nonconformity: Evidence from home videos. *Developmental Psychology*, *44*, 46–58.

Rieger, G., & Savin-Williams, R. C. (2012). Gender nonconformity, sexual orientation, and psychological well-being. *Archives of Sexual Behavior*, *41*, 611–621. http://doi.org/10.1007/s10508-011-9738-0

Roberts, A. L., Rosario, M., Corliss, H. L., Koenen, K. C., & Austin, S. B. (2012). Childhood gender nonconformity: A risk indicator for childhood abuse and posttraumatic stress in youth. *Pediatrics*, *129*, 410–417. http://doi.org/10.1542/peds.2011-1804

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

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*, 347–363. http://doi.org/10.1177/0734282911406661

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, *17*, 241–264. http://doi.org/10.1080/10705511003659375

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, *3*, 153–160. http://doi.org/10.1.1.18.3086

Singh, D., Deogracias, J. J., Johnson, L. L., Bradley, S. J., Kibblewhite, S. J., Owen-Anderson, A., … Zucker, K. J. (2010). The Gender Identity/Gender Dysphoria Questionnaire for Adolescents and Adults: Further validity evidence. *Journal of Sex Research*, *47*, 49–58. http://doi.org/10.1080/00224490902898728

Singh, D., McMain, S., & Zucker, K. J. (2011). Gender identity and sexual orientation in women with borderline personality disorder. *Journal of Sexual Medicine*, *8*, 447–454. http://doi.org/10.1111/j.1743-6109.2010.02086.x

Skidmore, W. C., Linsenmeier, J. A. W., & Bailey, J. M. (2006). Gender nonconformity and psychological distress in lesbians and gay men. *Archives of Sexual Behavior*, *35*, 685–697. http://doi.org/10.1007/s10508-006-9108-5

Steensma, T. D., Biemond, R., de Boer, F., & Cohen-Kettenis, P. T. (2011). Desisting and persisting gender dysphoria after childhood: A qualitative follow-up study. *Clinical Child Psychology and Psychiatry*, *16*, 499 –516. http://doi.org/10.1177/1359104510378303

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, *9*, 1–12. http://doi.org/10.1027/1614-2241/a000049

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. http://doi.org/10.1177/109442810031002

Veale, J. F. (2011). *Biological and psychosocial correlates of gender-variant and gender-typical identities.* Unpublished doctoral thesis, Massey University, Auckland, New Zealand.

Veale, J. F., Clarke, D. E., & Lomax, T. C. (2008). Sexuality of male-to-female transsexuals. *Archives of Sexual Behavior*, *36*, 586–597. http://doi.org/10.1007/s10508-007-9306-9

Veale, J. F., Clarke, D. E., & Lomax, T. C. (2010). Biological and psychosocial correlates of adult gender-variant identities: New findings. *Personality and Individual Differences*, *49*, 252–257. http://doi.org/10.1016/j.paid.2010.03.045

Weinrich, J. D., Atkinson, J. H., Jr., McCutchan, J. A., & Grant, I. (1995). Is gender dysphoria dysphoric? Elevated depression and anxiety in gender dysphoric and nondysphoric homosexual and bisexual men in an HIV sample. HNRC Group. *Archives of Sexual Behavior*, *24*, 55–72.

Williams, J. S. (1978). A definition for the common factor model and the elimination of problems of factor score indeterminacy. *Psychometrika*, *43*, 293–306. http://doi.org/10.1007/BF02293640

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. http://doi.org/10.1111/0081-1750.00078

Zucker, K. J., Bradley, S. J., Owen-Anderson, A., Kibblewhite, S. J., Wood, H., Singh, D., & Choi, K. (2012). Demographics, behavior problems, and psychosexual characteristics of adolescents with gender identity disorder or transvestic fetishism. *Journal of Sex & Marital Therapy*, *38*, 151–189. http://doi.org/10.1080/0092623X.2011.611219

Zucker, K. J., Mitchell, J. N., Bradley, S. J., Tkachuk, J., Cantor, J. M., & Allin, S. (2006). The Recalled Childhood Gender Identity/Gender Role Questionnaire: Psychometric properties. *Sex Roles*, *54*, 469–483. http://doi.org/10.1007/s11199-006-9019-x

Footnotes

[1] Readers interested in this type of analysis are referred to Sass (2011) and Vandenberg and Lance (2000) for articles that provide an overview of this analysis.

[2] This corresponds to a standardized factor loading of .70 or less. The *cosmetics* item was only administered to birth-assigned males.

[3] I thank anonymous reviewers for alerting me to these points.