Supplemental Material For:
Predicting Early Childhood Gender Transitions

This document contains (a) details about methods and results not reported in the main text and (b) results for analyses not reported in the main text.

Section 1: Details on Methodology and Analyses from Main Text

Participants

Families of transgender and gender nonconforming children heard about this research via at least one of the following: a conference/camp for gender nonconforming and transgender children, a support group for families of gender nonconforming and transgender children, a media story about past work by this research team, someone else who had participated in the study, or someone who had heard about the study, through an Internet search, or through the same university database controls came from. Participation occurred at children's homes or privately in a public location (e.g., a meeting room of a local library). Control participants were recruited from a university database of families interested in research on child development in the Pacific Northwest and were tested in a research laboratory. All participants lived in the United States or Canada.

Measures and Data Preparation

The present analyses focused on a composite made of five measures of gender development. In addition to these measures, all children received some other measures, but those measures differed by age of child, explored unrelated constructs (e.g., anxiety), or were only given to a subset of participants. Thus, these measures are not reported in either here or in the main text.

Missing Data

We obtained 20 imputed datasets using the mice package (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2016). Specifically, we used predictive mean matching and used 20 iterations for each imputation. Raghunathan and colleagues (2002) suggested that 10 iterations should be sufficient for most purposes. We examined trace plots to check for convergence.

Bayesian Estimation Approach

We used brms (Bürkner, 2016) defaults by generating four Markov chain Monte Carlo (MCMC) chains composed of 2,000 total iterations (1,000 warm-up). However, as we conducted our analyses on 20 imputed datasets (see above), we obtained 80 total chains (4 chains per dataset × 20 datasets). Pooling across these chains was executed using the brms function "brm_multiple". We numerically assessed chain vergence by examining the Gelman-Rubin statistic $\hat{R}$ (Gelman & Hill, 2006), and in all cases, we found evidence that our chains converged (i.e., $\hat{R} < 1.05$). We also visually assessed convergence by examining trace plots. Specifically, we looked for two properties in our trace plots (McElreath, 2016): *stationarity* and *good mixing*. For

all model parameters, all chains appeared to converge such that they demonstrated these two properties. Finally, we used posterior predictive checks (Gelman & Hill, 2006) to look for obvious sources of model misfit. We generally found that data simulated under our models was similar to the data we observed.

Priors for Analysis 1: Do Gender Identity and Preferences Predict Social Transitions? We used logistic regression models to investigate this research question. Gelman and colleagues (2008) recommended using Cauchy priors (i.e., a Student-*t* with 1 degree of freedom) with scale 2.5 and 10 for logistic regression coefficients and intercepts, respectively. As thick tails (e.g., a Student-*t* with 1 degree of freedom) can in some instances be problematic (e.g., see Ghosh, Li, & Mitra, 2018), we instead used Student-*t* priors with 4 degrees of freedom (which has thinner tails than a Student-*t* with 1 degree of freedom) for regression coefficients (M=0, SD=2.5) and intercepts (M=0, SD=10) reported in the main text. These priors are weakly informative insofar that they contain minimal information, yet help yield stable and accurate inferences (e.g., Gelman et al., 2008). To illustrate, a change of 2.5 (the *SD* for our Student-*t* prior on regression coefficients) in logistic regression moves a probability from .08 to .50, or from .50 to .92. Thus, it seems reasonable to select a prior distribution that assigns most of the weight to values of less than 2.5. Because we wanted our regression intercept to be able to take on a very wide range of values, we chose an even wider prior distribution (SD=10) for regression intercepts.

Priors for Analysis 2: Do Gender Identity and Preferences Differ between Future Transitioners, Transgender Children, and Controls? We used a multilevel beta regression to investigate this research question. Specifically, with gender identity and preference scores as the outcome variable, our first model included only group (future transitioners vs. transgender vs. control) as a random effect. For this model, we used the same prior as Analysis 1 for the intercept (Student-*t* with 4 degrees of freedom, M=0, and SD = 10). For the variance component of the model, we used more informative priors because of the small number of groups (three). In cases in which the number of groups is small, priors that allow extreme values can be problematic and cause divergent transitions (McElreath, 2016, p. 363-364). Indeed, we similarly found that using wide priors on the variance component, such as a half Student-*t* (df=4, M=0, SD=2.5), resulted in divergent transitions. Thus, for our first model we elected to use a prior with a smaller *SD* on the variance component by using a gamma prior (*shape*=2.5, *rate*=7.5). Some of the divergent transitions seemed to have occurred because the standard deviation of the group-level parameters was getting too close to zero. We chose the shape of our gamma prior to shift the posterior mode away from zero (though still permitting it to be very close to zero). We also increased the "adapt_delta" argument to .999 and changed the "max_treedepth" argument to 15. Our second model added covariates as fixed effects and paralleling our strategy for Analysis 1 (as we also used a logit link), we used Student-*t* priors (df=4, M=0, SD=2.5) for regression coefficients. We attempted to model factor variables with two levels (e.g., participant sex) as random effects but encountered divergent transitions using a variety of priors.

Multiverse Analyses

We examined the sensitivity of our results to 3 data processing decisions: (1) method of combining our five gender development measures (e.g., using different combinations of four of the five measures, three of the five measures, etc.); (2) missing data approach (ignore missingness vs. multiple imputation); and (3) decision to remove influential or unusual observations (ignore vs. exclude cases).

To identify influential cases, we used the cross-validated leave-one-out (LOO) predictive distribution (as recommended by Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2017). Specifically, we used the loo package (Vehtari, Gelman, & Gabry, 2017a), which "… automatically computes an empirical estimate of how similar the full-data predictive distribution is to the LOO predictive distribution for each left out point" (Gabry et al., 2017, p. 12). The loo package returns the diagnostic $\hat{k}$ from comparing the LOO predictive distribution and the full-data predictive distribution, and a large $\hat{k}$ value indicates that a data point is highly influential. As for choosing what values of are $\hat{k}$ are "large", Vehtari, Gelman, and Gabry (2017b) observed good performance for $\hat{k}$ values of less than .7. Moreover, .7 is the default threshold in the loo package for identifying influential observations (i.e., data points with $\hat{k} > .7$ are flagged). Thus, we anticipated refitting models for our multiverse analyses after removing data points with $\hat{k} > .7$. However, we observed no such cases when conducting our analyses.

Section 2: Details on Analyses Reported only in Supplemental Material

Model Comparison

In addition to our Bayesian estimation approach, we again used the loo package (Vehtari, Gelman, & Gabry, 2017a) to compare different models using (1) leave-one-out cross-validation (LOO-CV; Vehtari, Gelman, & Gabry, 2017b); models weights based on (2) bootstrapped-Pseudo-Bayesian Model Averaging (hereafter, Pseudo-BMA+) and (3) Bayesian stacking (Yao, Vehtari, Simpson, & Gelman, 2018). LOO-CV estimates how well a model would predict observations in a new sample. We used LOO-CV to assess whether more complex models had greater out-of-sample prediction accuracy relative to more simplistic models. Similarly, Pseudo-BMA+ involves (1) separately estimating the predictive performance of candidate models and (2) constructing weights (that sum to 1.0) for each model based on its *relative* predictive performance. Bayesian stacking jointly estimates model weights with the purpose of finding the best weights (again, that sum to 1.0) for combining predictive distributions. For useful discussion and examples comparing Pseudo-BMA+ and Bayesian stacking weights, see here: http://mc-stan.org/loo/articles/loo2-weights.html.

For each analysis, we compared four models. For Analysis 1, we fit the following models:

- Model 1: a model with no predictors (i.e., an intercept only model)
- Model 2: a model with our gender identity and preference measure as the sole predictor
- Model 3: a model with only covariates (e.g., age, sex, etc.) as predictors
- Model 4: a model with covariates and our gender identity and preference measure as predictor variables

For Analysis 2, we fit the following models:

- Model 1: a model with no predictors and a fixed intercept
- Model 2: a multilevel model that included a unique intercept for each group of participants (i.e., varying intercept model)
- Model 3: a model with only covariates (e.g., age, sex, etc.) as predictors and a fixed intercept
- Model 4: a model with covariates and a unique intercept for each group of participants

The models fit for Analysis 3 were the same as those fit for Analysis 1.

Sensitivity Analysis Using Different Priors

Analysis 1. To test the sensitivity of our results to different priors on regression coefficients (results did not meaningfully change by altering the prior for the intercept), we also examined our results after altering the *SD* of the Student-*t* prior placed on regression coefficients to (a) *SD*=5 (less informative) and (b) *SD*=1 (more informative). These results are reported below.

Analysis 2. To test the sensitivity of our results to using more/less informative priors, we also refit our initial model (containing only group as a random effect) after altering the *rate* for the gamma prior on the variance component of the model. Specifically, we changed the *rate* to: (a) *rate*=5 (less informative) and (b) *rate*=10 (more informative). For our second model that included covariates as fixed effects, we similarly adapted the *SD* of the Student-*t* prior placed on regression coefficients to (a) *SD*=5 and (b) *SD*=1 (i.e., the same priors used in our sensitivity analysis for Analysis 1).

Priors for Analysis 3: Do Gender Identity and Preferences Differ between Non-Transitioners, Transgender Children, and Controls?

As this research question was parallel to that asked in Analysis 2 (except we focused on non-transitioners instead of future transitioners), we used the same priors as those used in Analysis 2 (including a sensitivity analysis).

Multiverse Analysis for Analysis 3

We used the same criteria for identifying unusual/influential cases as used in the multiverse analyses for Analysis 1 and Analysis 2. We observed no instances in which $\hat{k} > .7$.

Section 3: Descriptive Statistics and Zero-Order Correlations

Table S1. Average scores on the five gender cognition measures and the gender cognition composite by group.

Variable	Gender Nonconforming		Transgender		Control	
	Future Transitioners (n = 36)	Non-Transitioners (n = 49)	Matched to Future Transitioners (n = 35)	Matched to Non-Transitioners (n = 49)	Matched to Future Transitioners (n = 36)	Matched to Non-Transitioners (n = 49)
Peer (*M*, *SD*)	0.77 (0.28)	0.67 (0.26)	0.79 (0.22)	0.77 (0.23)	0.79 (0.20)	0.79 (0.25)
Toy (*M*, *SD*)	0.64 (0.24)	0.58 (0.24)	0.65 (0.23)	0.66 (0.21)	0.68 (0.20)	0.68 (0.21)
Clothing (*M*, *SD*)	0.81 (0.22)	0.68 (0.33)	0.88 (0.16)	0.85 (0.19)	0.87 (0.14)	0.84 (0.16)
Similarity (*M*, *SD*)	0.66 (0.18)	0.60 (0.19)	0.74 (0.17)	0.73 (0.18)	0.69 (0.17)	0.71 (0.17)
Identity (*M*, *SD*)	0.76 (0.22)	0.54 (0.33)	0.86 (0.21)	0.88 (0.18)	0.79 (0.21)	0.85 (0.20)
Composite (*M*, *SD*)	0.73 (0.16)	0.61 (0.21)	0.78 (0.14)	0.78 (0.12)	0.76 (0.11)	0.77 (0.12)

Note: All variables took on values between 0 and 1. For missing values, we used the average score across the 20 imputed datasets.

Table S2. Zero-order correlations among gender cognition scores and demographic variables.

		1	2	3	4	5
1	Gender Cognition (Range: 0,1)	–				
2	Sex (0=Female; 1=Male)	-0.09	–			
3	Age (months)	-0.07	-0.01	–		
4	Race (0=Non-White; 1=White)	0.01	-0.06	0.02	–	
5	Parent Political Orientation (Range: 1,7)	-0.11	0.11	-0.04	0.07	–
6	Parent Income (Range: 1,5)	0.15	-0.05	0.03	-0.07	-0.02

Note: df= 254 for all correlations. For missing values, we used the average score across the 20 imputed datasets.

Table S3. Zero-order correlations among variables used in logistic regression analyses testing the association between gender nonconformity and social transition status (i.e., Analysis 1). All children in this analysis were gender nonconforming (future transitioners and non-transitioners).

		1	2	3	4	5	6	7
1	Transition (0=No; 1=Yes)	–						
2	Gender Cognition (Range: 0,1)	0.29	–					
3	Sex (0=Female; 1=Male)	0.24	-0.07	–				
4	Age (months)	-0.20	-0.01	0.01	–			
5	Race (0=Non-White; 1=White)	0.10	0.10	0.08	-0.08	–		
6	Time Since Initial Test (months)	0.24	0.08	0.02	-0.01	0.15	–	
7	Parent Political Orientation (Range: 1,7)	-0.12	-0.19	-0.14	0.05	-0.03	0.05	–
8	Parent Income (Range: 1,5)	0.00	0.11	0.01	0.09	-0.06	0.16	0.01

Note: df= 85 for all correlations. For missing values, we used the average score across the 20 imputed datasets.

Table S4. Zero-order correlations among variables (excluding group – treated as a random effect in our analyses) used in multilevel beta regression analyses testing whether gender cognition scores varied between *future transitioners* and their covariate-matched transgender and control peers (i.e., Analysis 2).

		1	2	3	4	5
1	Gender Cognition (Range: 0,1)	–				
2	Sex (0=Female; 1=Male)	-0.09	–			
3	Age (months)	-0.30	0.02	–		
4	Race (0=Non-White; 1=White)	0.01	-0.16	0.10	–	
5	Parent Political Orientation (Range: 1,7)	-0.20	0.20	0.02	-0.09	–
6	Parent Income (Range: 1,5)	0.20	-0.09	-0.06	-0.06	-0.11

Note: df= 107 for all correlations. For missing values, we used the average score across the 20 imputed datasets.

Table S5. Zero-order correlations among variables (excluding group – treated as a random effect in our analyses) used in multilevel beta regression analyses testing whether gender cognition scores varied between *non-transitioners* and their covariate-matched transgender and control peers (i.e., Analysis 3). This analysis was not reported in the main text.

		1	2	3	4	5
1	Gender Cognition (Range: 0,1)	–				
2	Sex (0=Female; 1=Male)	-0.10	–			
3	Age (months)	0.09	-0.01	–		
4	Race (0=Non-White; 1=White)	0.02	0.02	-0.05	–	
5	Parent Political Orientation (Range: 1,7)	-0.07	0.04	-0.06	0.19	–
6	Parent Income (Range: 1,5)	0.13	-0.03	0.09	-0.07	0.04

Note: df= 147 for all correlations. For missing values, we used the average score across the 20 imputed datasets.

Section 4: Full Results for Models Estimated in Analyses 1-3

The following tables present estimates for all predictor variables included in the models for Analyses 1-3.

Table S6. Results of the logistic regression analyses examining the extent to which gender cognition scores predict social transition status both with (Model 1) and without (Model 2) controlling for demographic covariates. These models were fit for Analysis 1.

Model	Variable	Coefficient	Coefficient 95% HDI	OR	OR 95% HDI
Model 1					
	Gender Cognition	1.44	[0.44, 2.50]	4.22	[1.55, 12.20]
Model 2					
	Gender Cognition	1.65	[0.47, 2.84]	5.20	[1.60, 17.11]
	Age	-0.84	[-1.93, 0.22]	0.43	[0.15, 1.24]
	Parent Income	-0.32	[-1.38, 0.69]	0.72	[0.25, 2.00]
	Time Since Initial Test	1.25	[0.13, 2.40]	3.51	[1.14, 11.01]
	Parent Political Orientation	-0.07	[-1.10, 0.98]	0.93	[0.33, 2.67]
	Race	0.11	[-1.01, 1.22]	1.11	[0.36, 3.40]
	Sex	1.45	[0.29, 2.65]	4.26	[1.34, 14.13]

Note: OR = odds ratio. HDI = highest density interval.

Table S7. Results of the multilevel beta regression analyses testing whether gender cognition scores varied between *future transitioners* and their covariate-matched transgender and control peers. Estimates are presented both with (Model 1) and without (Model 2) controlling for demographic covariates. These models were fit for Analysis 2.

Model	Variable Type	Variable	Estimate	Estimate 95% HDI	OR	OR 95% HDI
Model 1						
	Random effect					
		Group (SD)	0.19	[0.02, 0.50]	1.19	[1.00, 2.12]
	Fixed effect					
		Intercept	1.14	[0.78, 1.48]	3.13	[2.18, 4.39]
Model 2						
	Random effect					
		Group (SD)	0.25	[0.03, 0.58]	1.28	[1.03, 1.79]
	Fixed effect					
		Intercept	1.18	[0.78, 1.58]	3.25	[1.76, 5.98]
		Age	-0.41	[-0.70, -0.13]	0.66	[0.50, 0.88]
		Parent Income	0.29	[-0.01, 0.57]	1.33	[0.99, 1.77]
		Parent Political Orientation	-0.49	[-0.82, -0.16]	0.61	[0.44, 0.85]
		Race	0.17	[-0.14, 0.48]	1.18	[0.87, 1.62]
		Sex	-0.07	[-0.46, 0.30]	0.93	[0.63, 1.34]

Note: OR = odds ratio. HDI=highest density interval

Table S8. Results of the multilevel beta regression analyses testing whether gender cognition scores varied between *non-transitioners* and their covariate-matched transgender and control peers. Estimates are presented both with (Model 1) and without (Model 2) controlling for demographic covariates. These models were fit for Analysis 3.

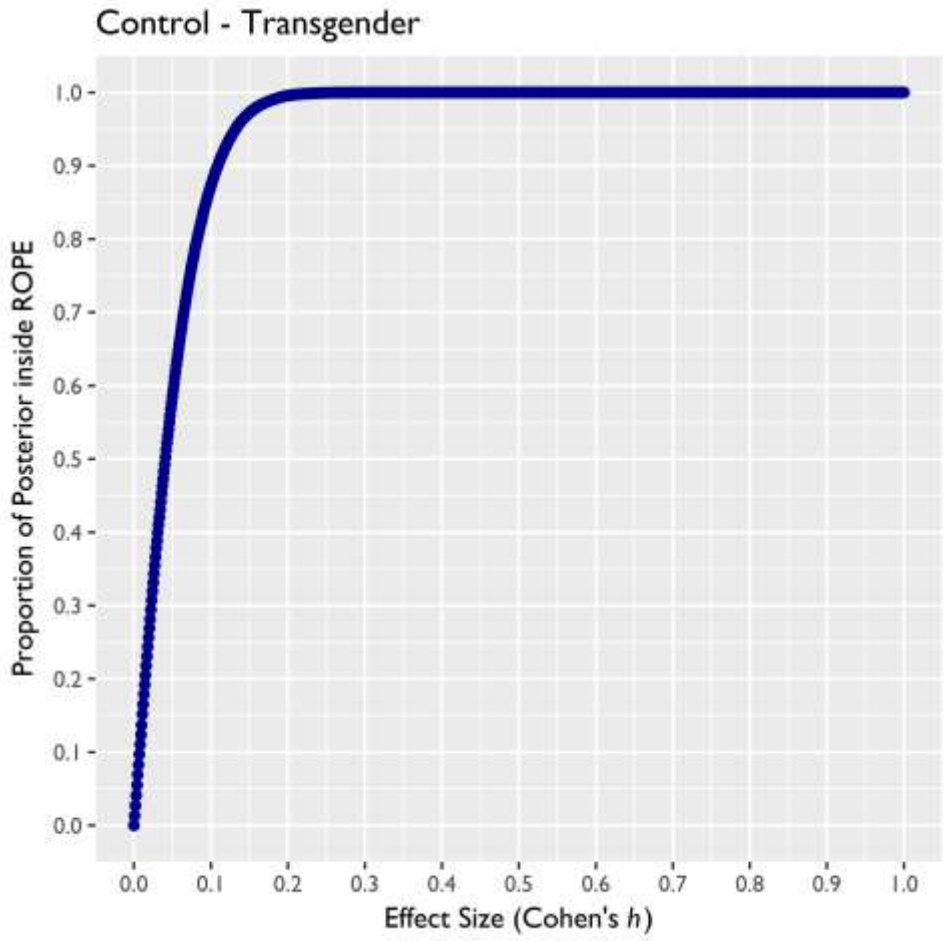
Model	Variable Type	Variable	Estimate	Estimate 95% HDI	OR	OR 95% HDI
Model 1						
	Random effect					
		Group (SD)	0.40	[0.16, 0.77]	1.72	[1.17, 2.16]
	Fixed effect					
		Intercept	0.91	[0.34, 1.49]	2.48	[1.40, 4.43]
Model 2						
	Random effect					
		Group (SD)	0.40	[0.15, 0.76]	1.72	[1.16, 2.14]
	Fixed effect					
		Intercept	0.93	[0.36, 1.48]	2.53	[1.43, 4.38]
		Age	0.25	[-0.03, 0.52]	1.28	[0.97, 1.68]
		Parent Income	0.08	[-0.20, 0.37]	1.09	[0.82, 1.45]
		Parent Political Orientation	0.07	[-0.27, 0.40]	1.07	[0.76, 1.50]
		Race	0.01	[-0.31, 0.33]	1.01	[0.73, 1.39]
		Sex	-0.16	[-0.45, 0.13]	0.85	[0.63, 1.14]

Note: OR = odds ratio. HDI=highest density interval

Section 5: Proportion of Posterior in ROPE (Analysis 2) as Function of Effect Size

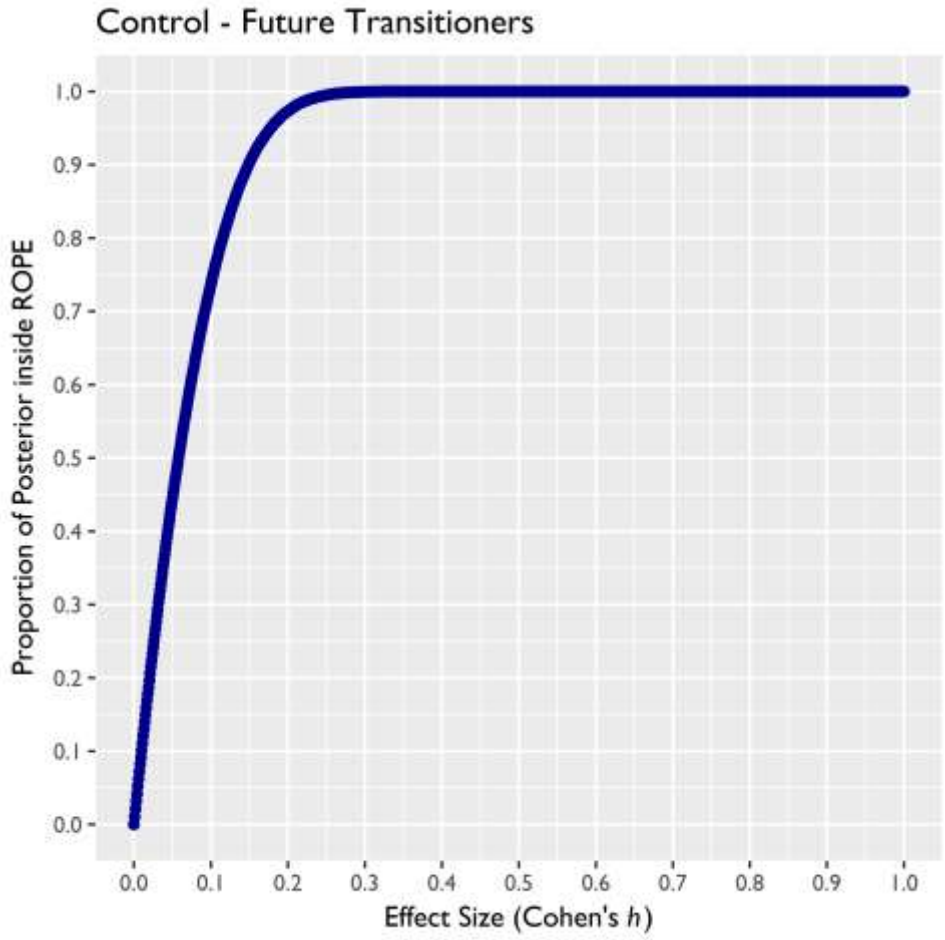
Without covariates

Figure S1. Proportion of posterior (for the difference in gender cognition scores between control and transgender participants) in the ROPE as a function of ROPE width (in effect size units).



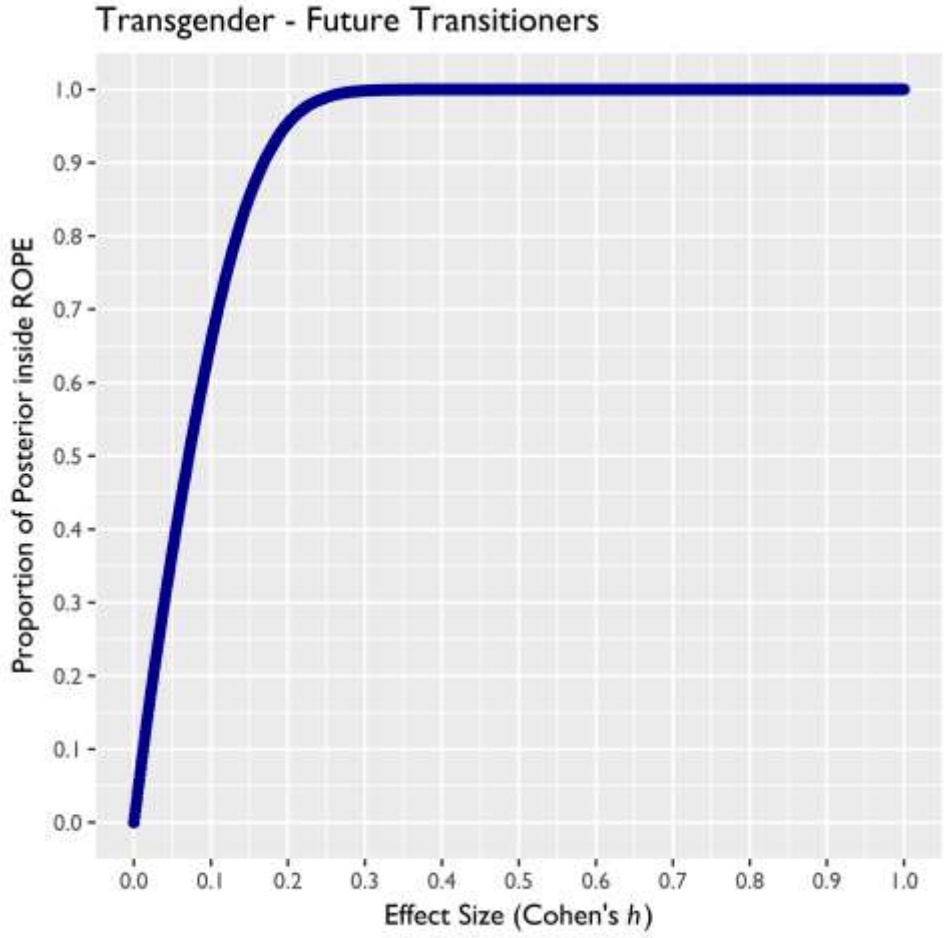
Note: Estimates came from model that excluded covariates.

Figure S2. Proportion of posterior (for the difference in gender cognition scores between control and future transitioners) in the ROPE as a function of ROPE width (in effect size units).



Note: Estimates came from model that excluded covariates.

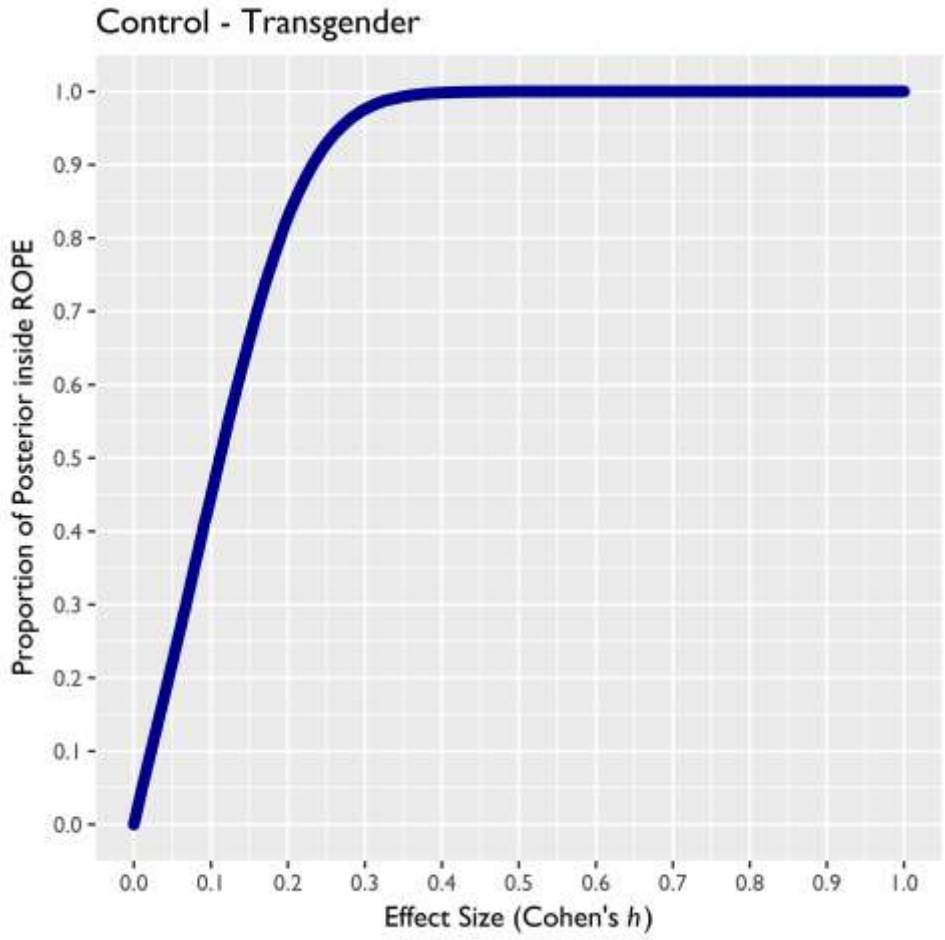
Figure S3. Proportion of posterior (for the difference in gender cognition scores between transgender and future transitioners) in the ROPE as a function of ROPE width (in effect size units).



Note: Estimates came from model that excluded covariates.

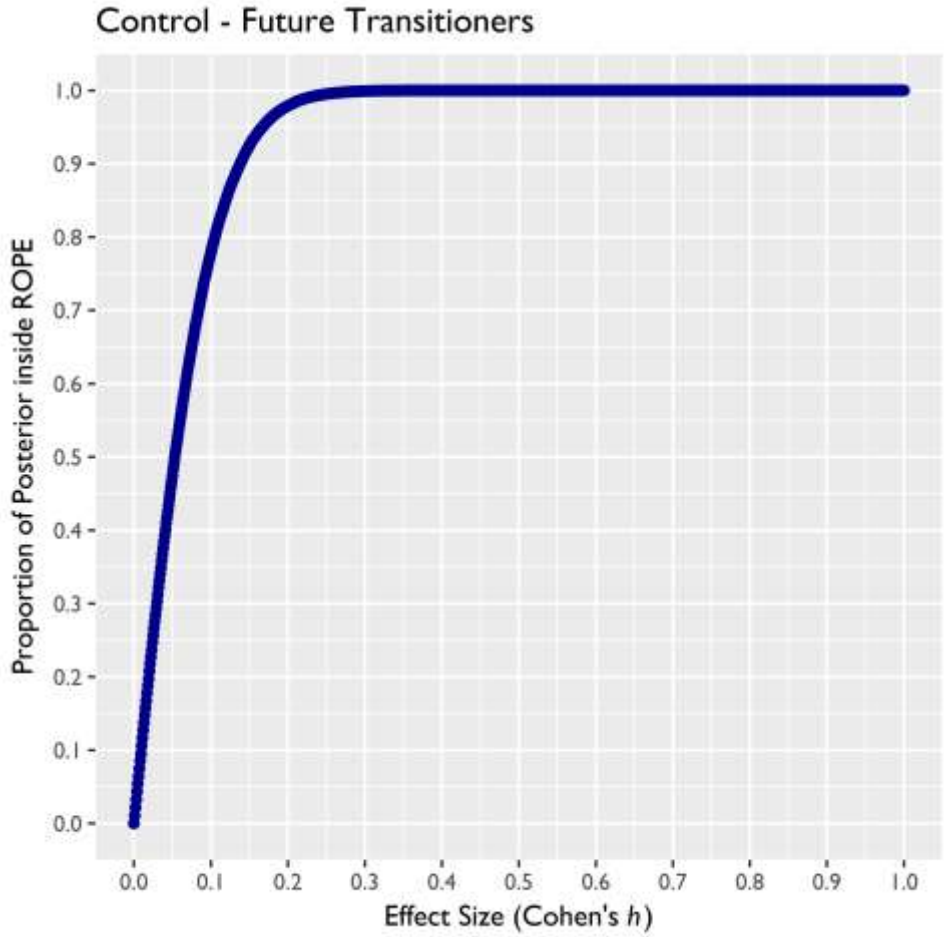
Including covariates

Figure S4. Proportion of posterior (for the difference in gender cognition scores between control and transgender participants) in the ROPE as a function of ROPE width (in effect size units).



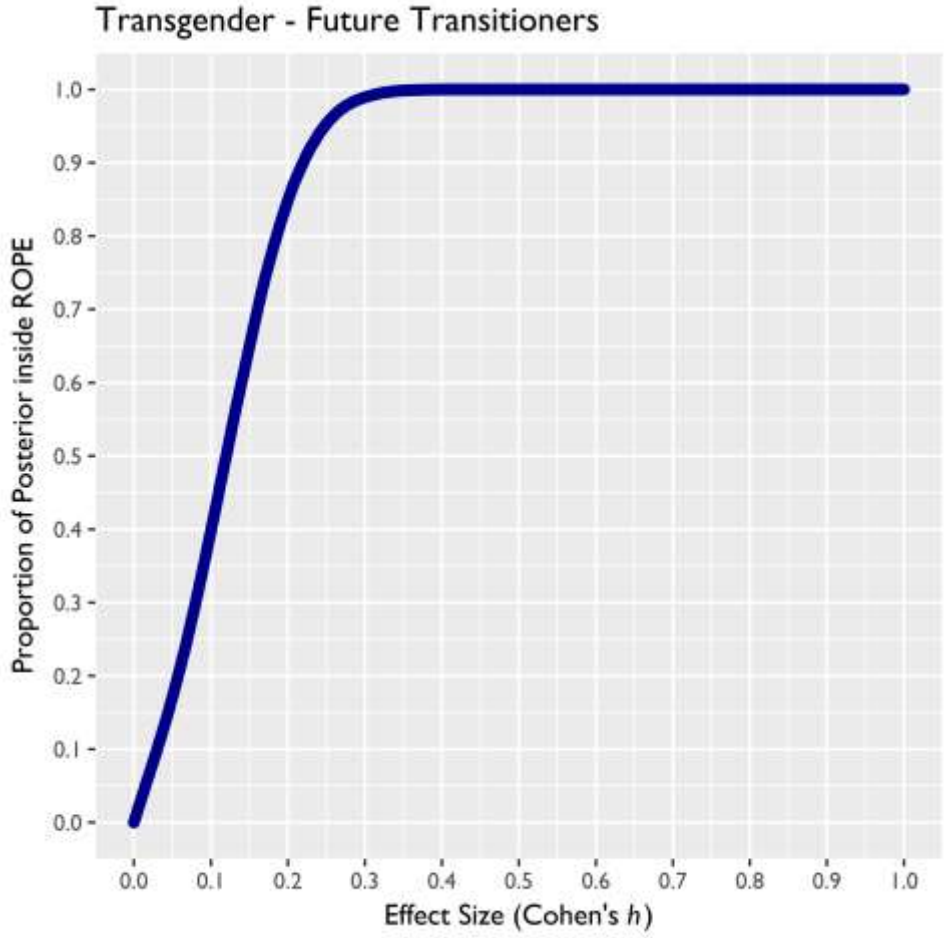
Note: Estimates came from model that included covariates.

Figure S5. Proportion of posterior (for the difference in gender cognition scores between control and future transitioners) in the ROPE as a function of ROPE width (in effect size units).



Note: Estimates came from model that included covariates.

Figure S6. Proportion of posterior (for the difference in gender cognition scores between transgender participants and future transitioners) in the ROPE as a function of ROPE width (in effect size units).



Note: Estimates came from model that included covariates.

Section 6: Model Comparison Results

The following tables present model comparison results (analyses in Supplemental Material only) for Analyses 1-3.

Table S9. LOO-CV comparisons between all models for Analysis 1 along with the SEs of the difference.

Description	LOO-CV	SE of Differences
Model 1 - Model 2	6.79	6.07
Model 1 - Model 3	1.72	7.63
Model 1 - Model 4	7.73	8.96
Model 2 - Model 3	-5.07	9.69
Model 2 - Model 4	0.94	7.14
Model 3 - Model 4	6.02	6.28

Note: A positive LOO-CV indicates that the second model improves predictive accuracy.

Table S10. Pseudo-BMA+ and Bayesian stacking weights for all models for Analysis 1.

	Pseudo-BMA+	Stacking
Model 1	0.07	0.00
Model 2	0.35	0.44
Model 3	0.13	0.18
Model 4	0.45	0.38

Results from Table S9 shows that models that contain the gender cognition measure (Models 2 and 4) are favored over an intercept only model (Model 1) or a model that only contains covariates (Model 3). Similarly, Table S10 shows that Models 2 and 4 received higher model weights than Models 1 and 3. In total, these results are consistent with our Bayesian estimation approach results which suggest that our gender cognition measure is a useful predictor of social transition status.

Table S11. LOO-CV comparisons between all models for Analysis 2 along with the SEs of the difference.

Description	LOO-CV	SE of Differences
Model 1 - Model 2	-0.99	1.86
Model 1 - Model 3	9.81	8.63
Model 1 - Model 4	12.02	9.26
Model 2 - Model 3	10.81	8.76
Model 2 - Model 4	13.01	8.86
Model 3 - Model 4	2.2	3.4

Note: A positive LOO-CV indicates that the second model improves predictive accuracy.

Table S12. Pseudo-BMA+ and Bayesian stacking weights for all models for Analysis 2.

	Pseudo-BMA+	Stacking
Model 1	0.07	0.18
Model 2	0.04	0.00
Model 3	0.26	0.00
Model 4	0.63	0.82

Results from Table S11 shows that a model with a random intercept for each group (Model 2) did not improve on the predictive accuracy of a single fixed intercept (Model 1). However, once covariates are included in the model, a model with both covariates and a random intercept for each group (Model 4) is preferred over a model with only covariates (Model 3) – which is also true when examining the model weights in Table S12. Taken together, there does seems to be systematic differences between groups after covariate adjustment. However, the ROPE analyses in the main-text and Figures S4-S6 shows that the credible effect sizes are "small" or "smaller than small".

Table S13. LOO-CV comparisons between all models for Analysis 3 along with the SEs of the difference.

Description	LOO-CV	SE of Differences
Model 1 - Model 2	21.76	10.39
Model 1 - Model 3	-4.52	6.99
Model 1 - Model 4	13.87	11.98
Model 2 - Model 3	-26.28	11.55
Model 2 - Model 3	-7.89	7.47
Model 3 - Model 4	18.39	8.97

Note: A positive LOO-CV indicates that the second model improves predictive accuracy.

Table S14. Pseudo-BMA+ and Bayesian stacking weights for all models for Analysis 3.

	Pseudo-BMA+	Stacking
Model 1	0.03	0.09
Model 2	0.81	0.90
Model 3	0.01	0.00
Model 4	0.15	0.01

Results from Table S13 shows that models that contain a unique intercept for group (Models 2 and 4) are favored over fixed intercept models (Models 1 and 3). Similarly, Table S14 shows that models containing a unique intercept for each group (but especially Model 2) had the largest model weights. Taken together, it appears that when comparing non-transitioners to covariate-matched transgender participants and controls, accounting for group membership is useful. Further, as shown below, most of the credible differences between groups is in the "small" to "medium" effect size range.

Section 7: Sensitivity Analysis to Different Prior Distributions

Table S15. Gender cognition estimates from models using more/less informative priors on regression coefficients for Analysis 1. Estimates for other variables (i.e., covariates) are not presented in Table S15.

Model	Prior on regression coefficients	Coefficient	Coefficient 95% HDI	OR	OR 95% HDI
Model 1					
	Student-*t* (*df*=4, *M*=0, *SD*=2.5)	**1.44**	**[0.44, 2.50]**	**4.22**	**[1.55, 12.20]**
	Student-*t* (*df*=4, *M*=0, *SD*=5.0)	1.50	[0.46, 2.57]	4.47	[1.59, 13.03]
	Student-*t* (*df*=4, *M*=0, *SD*=1.0)	1.22	[0.27, 2.21]	3.38	[1.31, 9.13]
Model 2					
	Student-*t* (*df*=4, *M*=0, *SD*=2.5)	**1.65**	**[0.47, 2.84]**	**5.20**	**[1.60, 17.11]**
	Student-*t* (*df*=4, *M*=0, *SD*=5.0)	1.76	[0.54, 3.03]	5.82	[1.72, 20.65]
	Student-*t* (*df*=4, *M*=0, *SD*=1.0)	1.30	[0.28, 2.42]	3.66	[1.33, 11.21]

Note: OR = odds ratio. HDI=highest density interval. Bolded results are those reported in the main-text. For all models, we placed a Student-*t* prior (*df*=4, *M*=0, *SD*=10) on the regression intercept.

Table S16. Gender cognition estimates (on a 0-1 scale) by group from models using more/less informative priors on the variance component of the multilevel beta regression model (Analysis 2).

		Control - Transgender		Control - Future Transitioners		Transgender - Future Transitioners	
Model	Prior on variance component	Estimate	Estimate 95% HDI	Estimate	Estimate 95% HDI	Estimate	Estimate 95% HDI
Model 1							
	gamma (*shape=2.5, rate=7.5*)	**-0.01**	**[-0.15, 0.11]**	**0.05**	**[-0.07, 0.20]**	**0.07**	**[-0.05, 0.22]**
	gamma (*shape=2.5, rate=5*)	-0.02	[-0.16, 0.11]	0.06	[-0.07, 0.21]	0.08	[-0.05, 0.23]
	gamma (*shape=2.5, rate=10*)	-0.01	[-0.14, 0.10]	0.05	[-0.07, 0.19]	0.06	[-0.05, 0.21]
Model 2							
	gamma (*shape=2.5, rate=7.5*)	**-0.11**	**[-0.29, 0.04]**	**0.01**	**[-0.16, 0.17]**	**0.12**	**[-0.02, 0.27]**
	gamma (*shape=2.5, rate=5*)	-0.13	[-0.32, 0.04]	0.00	[-0.18, 0.18]	0.13	[-0.01, 0.28]
	gamma (*shape=2.5, rate=10*)	-0.10	[-0.27, 0.05]	0.01	[-0.15, 0.16]	0.11	[-0.03, 0.25]

Note: HDI=highest density interval. Bolded results are those reported in the main-text. For all models, we placed a Student-*t* prior (*df*=4, *M*=0, *SD*=10) on the regression intercept. For Model 2, we also used a Student-*t* prior (*df*=4, *M*=0) and set the *SD* at 2.5, 5.0, and 1.0, respectively.

Table S17. Gender cognition estimates (on a 0-1 scale) by group from models using more/less informative priors on the variance component of the multilevel beta regression model (Analysis 3).

		Control - Transgender		Control - Future Transitioners		Transgender - Future Transitioners	
Model	Prior on variance component	Estimate	Estimate 95% HDI	Estimate	Estimate 95% HDI	Estimate	Estimate 95% HDI
Model 1							
	gamma (*shape=2.5, rate=7.5*)	**0.04**	**[-0.10, 0.18]**	**0.36**	**[0.20, 0.51]**	**0.32**	**[0.16, 0.47]**
	gamma (*shape=2.5, rate=5*)	0.04	[-0.10, 0.19]	0.37	[0.21, 0.53]	0.32	[0.17, 0.48]
	gamma (*shape=2.5, rate=10*)	0.04	[-0.10, 0.18]	0.35	[0.19, 0.51]	0.31	[0.15, 0.47]
Model 2							
	gamma (*shape=2.5, rate=7.5*)	**0.03**	**[-0.12, 0.18]**	**0.34**	**[0.17, 0.51]**	**0.31**	**[0.15, 0.48]**
	gamma (*shape=2.5, rate=5*)	0.03	[-0.12, 0.18]	0.35	[0.18, 0.53]	0.32	[0.16, 0.49]
	gamma (*shape=2.5, rate=10*)	0.03	[-0.12, 0.18]	0.33	[0.16, 0.50]	0.30	[0.14, 0.47]

Note: HDI=highest density interval. Bolded results are those reported in the main-text. For all models, we placed a Student-*t* prior (*df*=4, *M*=0, *SD*=10) on the regression intercept. For Model 2, we also used a Student-*t* prior (*df*=4, *M*=0) and set the *SD* at 2.5, 5.0, and 1.0, respectively.

Section 8: Results Omitting Participant with Short Follow-Up

One gender nonconforming participant had a much shorter period between their initial testing session and follow-up (3 months) compared to other participants. To see if we still found that gender preference and identity scores predicted socially transitioning after removing this unusual data point, we re-estimated the models reported in the main text after removing this participant from the dataset. Closely mirroring the results reported in the main text, our re-analysis found that participants expressing greater gender nonconformity in the initial testing session were more likely to socially-transition before follow-up both before, odds ratio=4.06, 95% HDI = [1.49–11.71], and after, odds ratio=5.01, 95% HDI = [1.60–17.07], controlling for covariates.

Section 9: Multiverse Analyses (Analysis 2)

Figure S7 displays the proportion of the posterior distribution of the ROPE for each of the 186 comparisons for Analysis 2. The take-away from Figure S7 is that (with few exceptions) most of the posterior distribution (often near 100%) was inside the ROPE for all between group-comparisons. Indeed, 162 of the 186 comparisons (87%) had more than 95% of the posterior distribution inside the ROPE. We also found that missing data approach had little impact on our results.

Section 10: Multiverse Analyses (Analysis 3)

Figure S8 presents the median differences between groups along with 95% HDIs of the differences (represented as the effect sizes) from our multiverse analysis comparing non-transitioners to matched control and transgender participants. Across the 186 comparisons (62 data sets × 3 between-group comparisons per data set), 106 comparisons had 95% HDIs that excluded zero (57% of comparisons) and only 47 of the comparisons fell completely inside the ROPE cutoffs (25% of comparisons). However, 95% HDIs for comparisons between control participants and transgender participants were never credibly different from zero (0/62=0%), whereas 85% of the comparisons between control participants and non-transitions (53/62) and transgender participants and non-transitioners (53/62) were credibly different from zero. Similarly, the proportion of comparisons in the ROPE varied as a function of whether the comparison was between non-transitioners and control participants (0/62=0%), control and transgender participants (47/62=76%), or non-transitioners and transgender participants (0/62=0%). That is, there were no instances in which comparing non-transitioners to matched transgender participants or control participants yielded a 95% difference that was completely consistent with a difference that is “small” or “smaller than small”. Further, Figure S8 also shows that he most plausible effect sizes for the difference are between a small (.20) and medium (.50) effect – especially for composite variables comprised of four or more gender development measures. Findings from Figure S9 mirror those from Figure S8 such that the proportion of the posterior distribution in the ROPE is low (i.e., often less than 50%) – especially as additional

gender development measures are added to the composite measure. Finally, we found that missing data approach had little impact on our results.

Figure S7. The proportion of the posterior distribution inside the ROPE cutoffs (0 ± Cohen's *h* of .20) for the multiverse analysis in which all gender development measures (31 columns) were used as outcomes in multilevel beta regression models with unique intercepts for each group using both missing data approaches (2 rows).

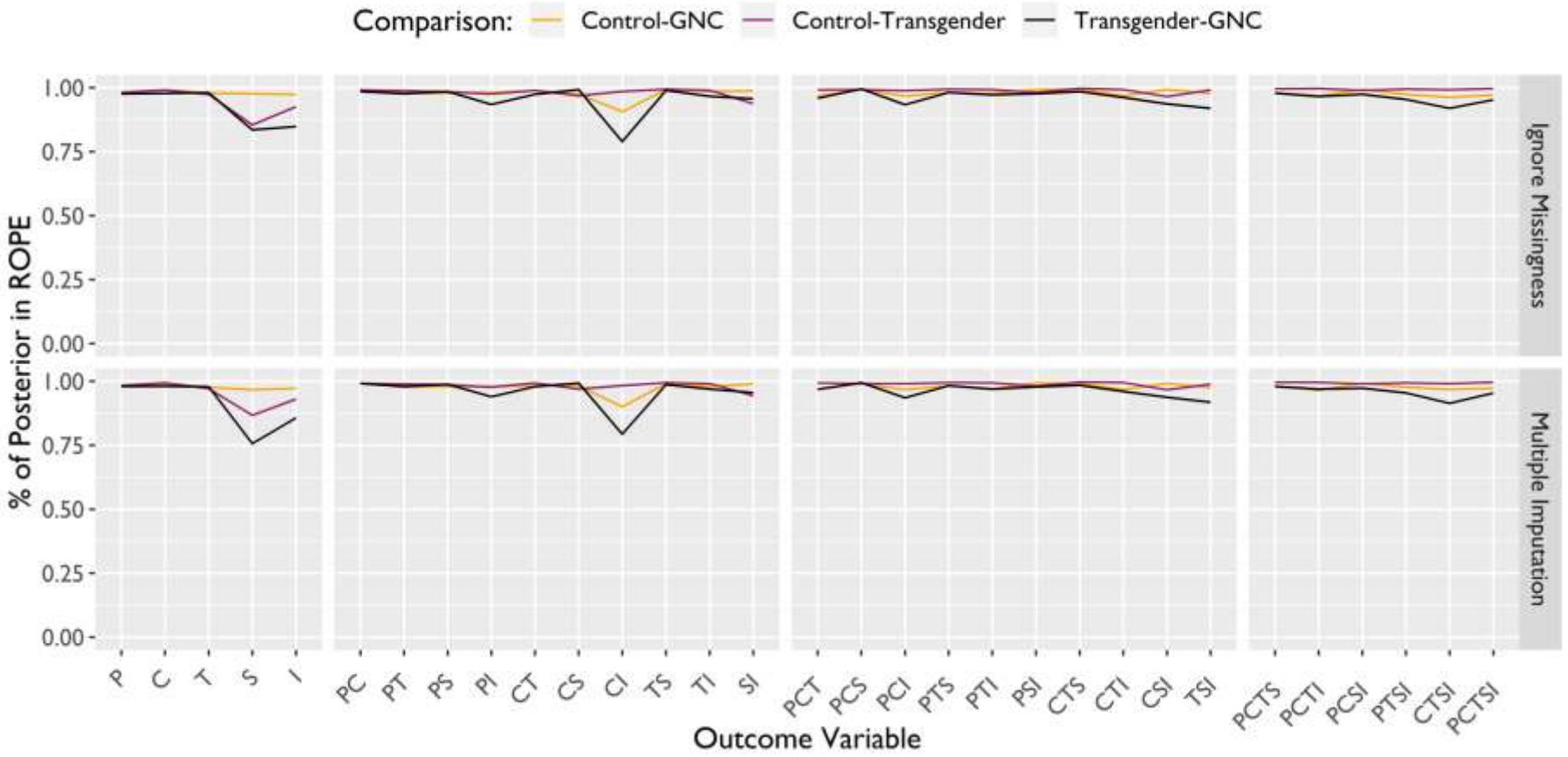


Figure S8. Multiverse analysis in which all gender development measures (31 columns) were used as outcomes in multilevel beta regression models with unique intercepts for each group using both missing data approaches (2 rows). Between-group differences were created and each estimate (dot) is the median between-group difference and intervals are 95% HDIs. Estimates and intervals are represented as effect size measures (Cohen's *h*). Dashed lines correspond to small negative (−.20) and small positive (.20) effect sizes, respectively. The ROPE is the area between these dashed lines. P=peer preferences, T=Toy preferences, C=Clothing Preferences, S=Gender Similarity, I=Gender Identity.

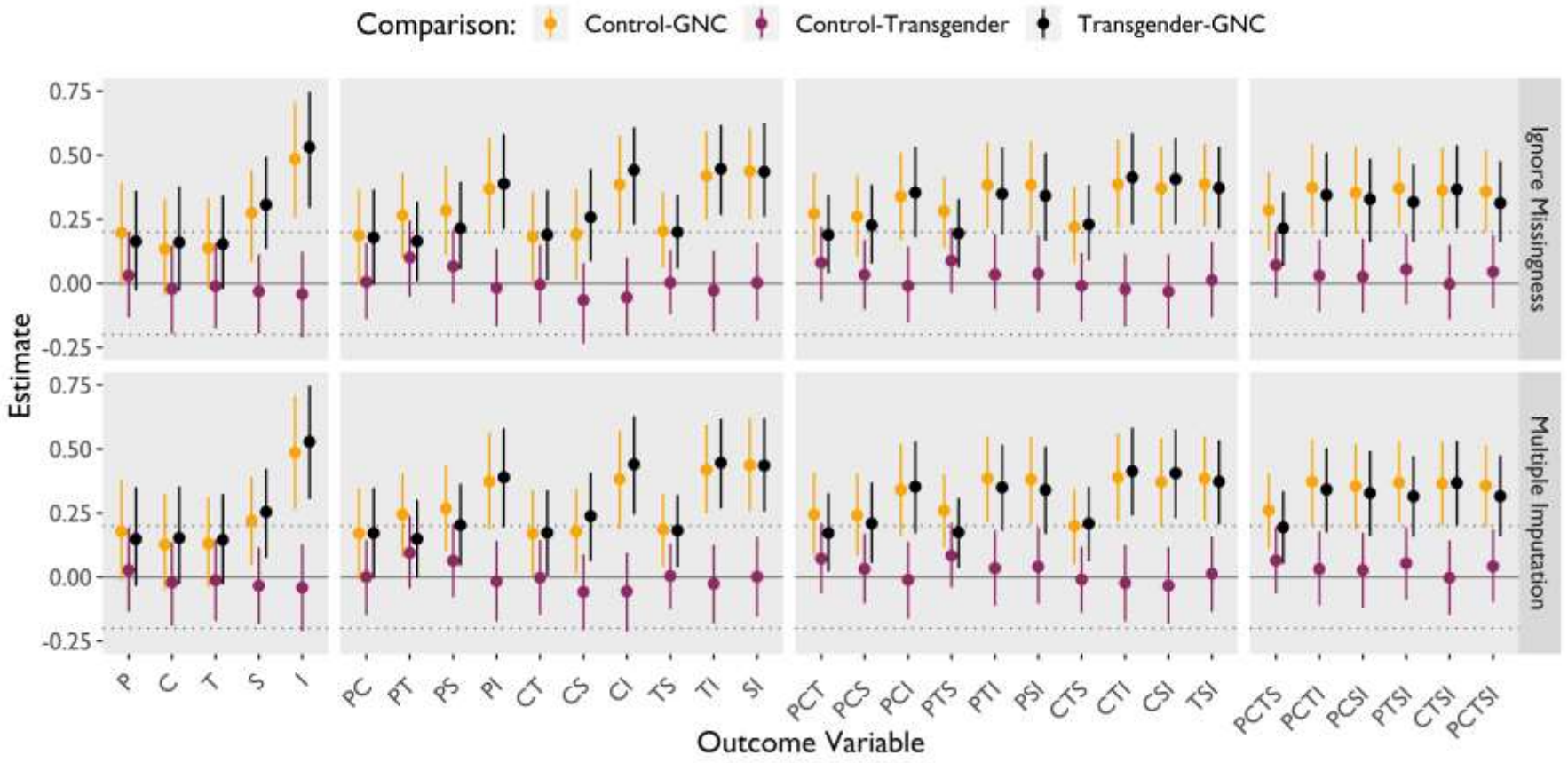
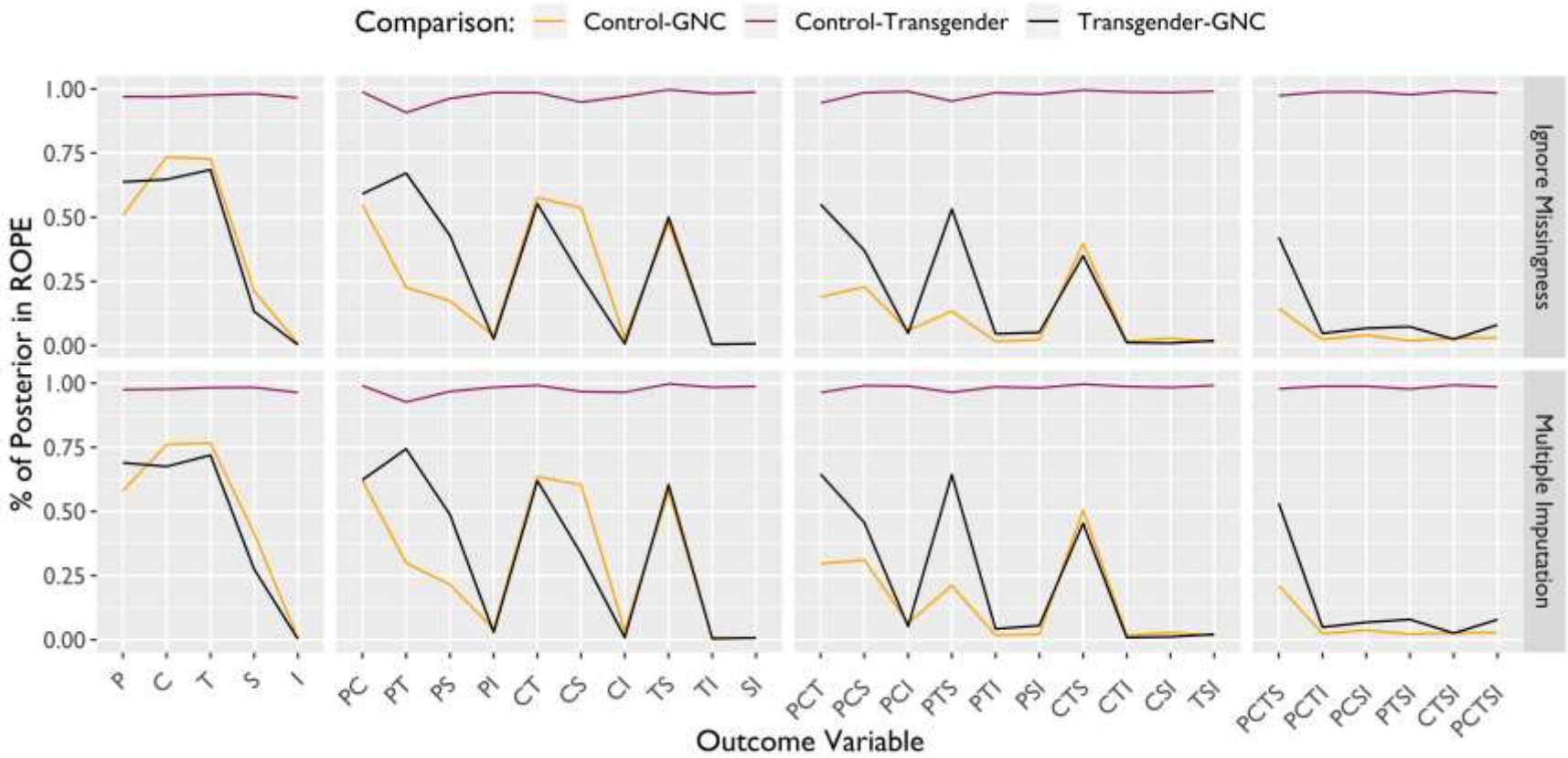


Figure S9. The proportion of the posterior distribution inside the ROPE cutoffs (0 ± Cohen's *h* of .20) for the multiverse analysis in which all gender development measures (31 columns) were used as outcomes in multilevel beta regression models with unique intercepts for each group using both missing data approaches (2 rows).



References

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80(*1).

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360-1383.

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*, 359-383.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: CRC Press

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raghunathan, T. E., Solenberger, P., & Van Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software installation instructions and user guide*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Van Buren, S., & Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1-67.

Vehtari, A., Gelman, A., & Gabry, J. (2017a). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 1.0.0, http://mc-stan.org/loo.

Vehtari, A., Gelman, A., & Gabry, J. (2017b). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413-1432.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*.