Infancy

WILEY

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES

BRIEF REPORT

The Infant Behavior Questionnaire Factor Structure Varies With Sample Characteristics

Anna M. Zhou[1,2] | Koraly Pérez-Edgar[3] | Vanessa LoBue[4] | Kristin A. Buss[3]

[1]Department of Psychiatry, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA | [2]Department of Psychology, University of Denver, Denver, Colorado, USA | [3]Department of Psychology, The Pennsylvania State University, University Park, Pennsylvania, USA | [4]Department of Psychology, Rutgers University-Newark, Newark, New Jersey, USA

Correspondence: Anna M. Zhou (anna.m.zhou@cuanschutz.edu)

Received: 13 February 2024 | **Revised:** 14 April 2025 | **Accepted:** 2 July 2025

Handling Editor: Eric Alf Walle

Funding: Data for this study is supported by grants from National Institutes of Health awarded to Pérez-Edgar, Buss, and LoBue (R01MH109692). Dr. Zhou was supported by a postdoctoral training grant (T32 MH015442). Drs. Buss and Pérez-Edgar's Psychology Professorships are supported by the Social Science Research Institute of The Pennsylvania State University, and endowments through the Tracy Winfree and Ted H. McCourtney Professorship in Children, Work, and Families (Buss) and the McCourtney Professorship of Child Studies (Pérez-Edgar).

Keywords: measurement invariance | negative affectivity | surgency | temperament

ABSTRACT

The Infant Behavior Questionnaire (IBQ) has been widely used to assess infant temperament traits, though there is limited empirical support for the recommended three-factor structure. The present study examined the replicability and measurement invariance in the IBQ using a large, multi-site longitudinal study of parent-child dyads (N = 357) in the United States. Temperament was reported by parents when infants were 4-, 8- and 12-months of age. Results show that the traditional three-factor structure did not fit our data well, and model modifications were needed to achieve acceptable fit. There was a high degree of covariance between the latent factors of surgency and orienting/regulation in modified three-factor models, suggesting that a modified two-factor model may be more appropriate for our data. Our findings also provide evidence that the modified three-factor structure is not invariant across sociodemographic groups. The findings highlight the need for researchers to examine the factor structure of the IBQ within their data before creating composites, especially in more diverse samples. If the three-factor structure does not replicate, we provide recommendations for alternate approaches to using the IBQ for developmental work.

1 | Introduction

Infant temperament has received considerable attention in developmental research as early temperament traits are associated with socioemotional adjustment and psychopathology risk (e.g., Ostlund et al. 2021; Rothbart and Bates 2006). Broadly conceptualized, temperament traits are biologically based individual differences in the domains of activity, reactivity, emotionality, and sociability (Shiner et al. 2012). While biologically based, contemporary views on temperament highlight the importance of experiential and contextual factors.

One widely accepted model of temperament focuses on individual differences in reactivity and self-regulation in the domains of activity, affect, and attention (Rothbart and Bates 2006). Reactivity includes arousability of emotional, motor, and attentional responses, and regulation includes behaviors that modulate reactivity (Gartstein and Rothbart 2003). To assess these behaviors, caregiver reports have the benefit of

tapping into the extensive knowledge base of caregivers, who have seen the child in many different contexts responding to a variety of naturally occurring stimuli (Rothbart and Bates 2006). The Infant Behavior Questionnaire (IBQ; Rothbart 1981; Gartstein and Rothbart 2003) is a widely used parent-report measure for infants aged 3 to 12 months, and assesses behaviors associated with fine-grained facets of temperament. From a factor analysis of a dataset of 3- to 12-month-old infants, Gartstein and Rothbart (2003) identified a higher order factor structure consisting of three broad dimensions—*Surgency/Extraversion*, *Negative Affectivity*, and *Orienting/Regulation*.

Despite its wide use, there is some evidence that the three-factor structure may not fit for all samples, especially from data collected from families with diverse cultural and sociodemographic backgrounds (e.g., Enlow et al. 2016). Moreover, prior work establishes that a consistent pattern of factor loadings can be identified only *after* model modification. To replicate standard loadings of the three-factor structure, researchers have allowed for cross-loadings of several subscales and the inclusion of error covariance terms (e.g., Gartstein et al. 2005; Enlow et al. 2016). These findings point to the lack of replicability in the three-factor structure, and that certain temperament traits may load onto multiple factors. Additionally, Enlow et al. (2016) found a lack of invariance on several factors and subscale scores by maternal country of birth, race/ethnicity, and household income. Thus, it may be that the canonical three-factor structure of the IBQ needs to be adjusted for use in diverse samples.

In this study, we examined the three-factor structure commonly used for the IBQ in 4-, 8- and 12-month-old infants from a multi-site longitudinal study. We also examined measurement invariance of the factor structure by infant sex, family income, and infant racial-ethnic majority/minority status.

2 | Methods

2.1 | Participants

Three hundred and fifty-seven infants (50.7% female) and parents were recruited from areas in and around University Park, Pennsylvania (Site 1), Harrisburg, Pennsylvania (Site 2), and Newark, New Jersey (Site 3) in the United States for a longitudinal study (Pérez-Edgar et al. 2021). Two hundred and ninety-eight infants were recruited when infants were 4 months of age (M_{age} = 4.80 months; SD_{age} = 0.80), with an additional 46 participants enrolled at 8 months (M_{age} = 8.83 months, SD_{age} = 0.73) and 13 participants enrolled at 12 months (M_{age} = 12.73, SD_{age} = 1.12). Demographic information of participants and their parents by site can be found in Table 1. The present study was conducted according to guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from a parent or guardian for each infant before any assessment or data collection. All procedures involving human subjects in this study were approved by the Institutional Review Boards at the Pennsylvania State University (IRB STUDY00004097) and Rutgers University (#Pro-2020000418). Parents provided written consent and were compensated for their participation. Our sample size was determined by the core research questions and aims of the parent grant and therefore was not specific to the current study.

2.2 | Procedures and Measures

2.2.1 | Questionnaires

Dimensions of infant temperament were assessed using caregivers' ratings on the Infant Behavior Questionnaire-Revised (IBQ-R; Gartstein and Rothbart 2003) at the 4, 8, and 12 month timepoints. Questionnaires were available in both English and Spanish and were administered based on the caregiver's first language. The IBQ-R is a 191-item survey designed to assess general patterns of behavior associated with temperament in infancy. Parents rated how often they observed a behavior in the past week. Each item describes an infant's behavior using a 7-point scale (1 = *never* to 7 = *always*). Parents are also given a "not applicable" response option when the infant has not been observed in the situation described. Each item loads onto 1 of 14 subscales: Activity Level, Distress to Limitations, Fear, Duration of Orienting, Smile/Laughter, High-intensity Pleasure, Low-intensity Pleasure, Soothability, Falling Reactivity, Cuddliness, Perceptual Sensitivity, Sadness, Approach, and Vocal Reactivity. The IBQ-R has demonstrated good internal consistency, reliability, and validity, including correlations with laboratory observations (Gartstein and Marmion 2008; Parade and Leerkes 2008).

2.3 | Data Analytic Plan

2.3.1 | Missing Data and Attrition

At 4 months, we had data from 274 infants (92% of the sample enrolled at 4 months) ranging from 2.76 to 6.90 months (M_{age} = 4.54, SD = 0.79). At 8 months, we had data from 240 infants (67% of the full sample) ranging from 6.41 to 10.71 months (M_{age} = 8.17, SD = 0.69). At 12 months, we had data from 214 infants (60% of the full sample) ranging from 11.14 to 16.16 months (M_{age} = 12.34, SD = 0.91). We examined associations between attrition and demographics (family income, parental education, infant sex, race/ethnicity, and full-term vs. preterm birth). Only maternal education was significantly associated with attrition at 8 ($\chi^2 = -0.12$, $p = 0.003$) and 12 months ($\chi^2 = -0.16$, $p < 0.001$). We used listwise deletion at each visit instead of missing data methods such as full-information maximum likelihood to most appropriately understand the structure of the data collected.

2.3.2 | Confirmatory Factor Analyses

Confirmatory factor analyses (CFA) were conducted in R 4.1.1 using the *lavaan* package (Rosseel 2012). Maximum likelihood estimation with robust standard errors were used. We first fit a three-factor solution to test the previously published model of the IBQ factor structure (Gartstein and Rothbart 2003). Model fit was evaluated using χ^2, the comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square

TABLE 1 | Sample demographics by Site assessed at enrollment (4-month), 8-month and 12-month.

Demographics	Site 1 University Park, PA *N* = 167	Site 2 Harrisburg, PA *N* = 81	Site 3 Newark, NJ *N* = 109	Combined *N* = 357
Infant sex				
Female	50.30%	45.68%	55.05%	50.70%
Male	49.70%	54.32%	44.95%	49.30%
Infant race				
Asian/Pacific Islander	4.35%	0%	2.13%	2.73%
Hispanic	4.96%	12%	46.81%	18.48%
White, non-Hispanic	81%	52%	8.51%	54.24%
Black, non-Hispanic	0%	22.67%	35.11%	15.15%
Native American	0%	0%	3.19%	0.91%
Multiracial	8.70%	13.33%	4.26%	8.48%
Maternal age at birth				
Mean (SD)	31.5 (4.42)	30.1 (4.53)	28.9 (6.14)	30.43 (5.09)
Family income at enrollment				
$15,000 or less	2.99%	16.05%	35.78%	15.97%
$16,000–20,000	2.40%	6.17%	11.93%	6.16%
$21,000–30,000	3.59%	7.41%	10.09%	6.44%
$31,000–40,000	4.79%	9.88%	3.67%	5.60%
$41,000–50,000	8.98%	3.70%	3.67%	6.16%
$51,000–60,000	13.77%	6.17%	1.83%	8.40%
Above $60,000	59.28%	41.98%	7.34%	39.50%
Decline to answer	4.19%	8.64%	25.69%	11.76%
Marital status at enrollment				
Married	87.12%	61.33%	44.68%	71.23%
Divorced	0.61%	1.33%	2.12%	1.05%
Single	4.91%	17.33%	27.66%	13.68%
Living with partner	7.36%	20%	26.60%	14.04%
Parental years of formal education at enrollment				
Parent 1 mean (SD)	17.38 (2.44)	15.15 (2.78)	13.28 (3.56)	15.92 (3.28)
Parent 2 mean (SD)	16.77 (2.74)	14.49 (2.80)	12.74 (3.50)	15.53 (3.31)
Questionnaire language at 8M				
English	100%	100%	55.74%	88.98%
Spanish	0%	0%	44.26%	11.02%
Questionnaire language at 12M				
English	100%	97.5%	60.34%	88.94%
Spanish	0%	2.5%	39.66%	11.06%

residual (SRMR), and the root mean square error of approximation (RMSEA). Good model fit is indicated by $p < 0.05$, CFI ≥ 0.95, TLI ≥ 0.95, RMSEA ≤ 0.08 (Hu and Bentler 1999), though fit indices of CFI ≥ 0.90 and TLI ≥ 0.90 may indicate acceptable fit. Modification indices were examined and included when our criteria for "good" model fit was not met. In the case of testing non-nested model structures, improvement in fit can be determined by ≥ 0.005 change in CFI, ≤ -0.010 change in RMSEA (Kass and Raftery 1995). We conducted post hoc power analyses using the *semPower* package (Moshagen and Bader 2024) to ensure that we had sufficient sample sizes to test our specified CFA models.

2.3.3 | Measurement Invariance

Measurement invariance based on infant sex, infant racial-ethnic majority/minority status, and family income were

tested through running configural, metric, and scalar invariance models where factor loadings are constrained. While we collected income data on all families, we did not have data on household size for a portion of our participants. On average, the household size was 4 based on data collected ($N = 157$). We then selected $30,000, the federal poverty level for a family of 4, as the criteria to test invariance by family income. Configural invariance means that the factor structure of the measure is the same across groups. Metric invariance refers to equal strengths of relations between subscales and their latent construct across groups. Lastly, scalar invariance means that individuals with the same scores on the latent variables would have the same scores on the observed items across groups; across-group differences in the means of the observed scale items are due to differences in the means of the underlying constructs.

3 | Results

3.1 | Descriptive Statistics

Descriptive statistics and Cronbach's alphas for subscales on the IBQ at 4, 8, and 12 months are in Table 2. Correlations among subscales at each time point are in Supporting Information S1 (Tables S1–S3).

3.2 | Confirmatory Factor Analyses

Model fit information for CFA models using 4-, 8- and 12-month data are presented in Table 3. The original three-factor solution did not have adequate model fit at any of the three timepoints. However, we were able to modify the original three-factor solution by allowing cross-loadings and covariances. Based on fit indices, the modified three-factor solutions fit the data significantly better than the original three-factor solution at all timepoints.

We then explored two-factor models given the high covariance between surgency and effortful control factors in the modified three-factor solutions. High covariance between factors suggests lack of orthogonality between these two factors. Given prior work that suggests that self-regulatory aspects to temperament modulate reactivity as the child develops (Rothbart 1989), we tested an alternative two-factor model focusing on reactivity. As this model was more exploratory, we allowed for items falling

TABLE 2 | Means, standard deviations, and Cronbach's alpha of IBQ subscales at 4, 8, and 12 months.

	4 months			8 months			12 months		
IBQ subscale	***M* (SD)**	**Range**	**α [95 CI]**	***M* (SD)**	**Range**	**α [95 CI]**	***M* (SD)**	**Range**	**α [95 CI]**
Activity level	4.04 (0.83)	1.93–6.27	0.79 [0.73, 0.83]	4.62 (0.72)	2.40–6.40	0.72 [0.65, 0.77]	4.57 (0.84)	1.50–6.33	0.75 [0.69, 0.80]
Distress to limitations	3.46 (0.77)	1.62–5.20	0.77 [0.70, 0.82]	3.69 (0.89)	1.56–6.13	0.84 [0.80, 0.87]	3.96 (0.86)	1.00–7.00	0.81 [0.75, 0.85]
Approach	4.57 (1.20)	1.00–7.00	0.85 [0.79, 0.89]	5.43 (0.83)	2.09–7.00	0.85 [0.80, 0.89]	5.69 (0.70)	3.00–7.00	0.80 [0.73, 0.85]
Fear	2.39 (1.00)	1.12–7.00	0.93 [0.91, 9.94]	2.68 (1.06)	1.00–6.12	0.94 [0.92, 0.95]	3.06 (1.06)	1.07–7.00	0.90 [0.88, 0.92]
Duration of orienting	4.20 (1.14)	1.25–7.00	0.87 [0.82, 0.90]	3.80 (1.05)	1.09–6.45	0.85 [0.79, 0.89]	3.80 (1.12)	1.00–6.33	0.87 [0.82, 0.90]
Smiling and laughter	4.75 (1.13)	2.11–7.00	0.84 [0.79, 0.88]	4.85 (1.04)	1.00–7.00	0.82 [0.77, 0.86]	5.10 (0.92)	1.00–7.00	0.80 [0.74, 0.84]
Vocal reactivity	4.37 (1.07)	1.33–7.00	0.82 [0.73, 0.88]	4.73 (0.98)	1.83–7.00	0.87 [0.82, 0.90]	5.30 (0.81)	3.00–7.00	0.83 [0.77, 0.86]
Sadness	3.18 (0.94)	1.00–7.00	0.81 [0.75, 0.86]	3.37 (0.97)	1.23–7.00	0.83 [0.78, 0.87]	3.34 (0.93)	1.00–6.00	0.75 [0.64, 0.82]
Perceptual sensitivity	3.70 (1.31)	1.00–7.00	0.88 [0.76, 0.93]	3.97 (1.28)	1.00–6.75	0.87 [0.81, 0.91]	4.42 (1.29)	1.00–7.00	0.90 [0.86, 0.93]
High intensity pleasure	5.35 (0.89)	3.00–7.00	0.81 [0.73, 0.86]	5.78 (0.70)	3.09–7.00	0.77 [0.68, 0.84]	5.99 (0.72)	1.00–7.00	0.83 [0.77, 0.87]
Low intensity pleasure	5.30 (0.87)	2.33–7.00	0.83 [0.78, 0.86]	5.04 (0.93)	2.00–7.00	0.89 [0.85, 0.92]	5.08 (0.89)	2.23–7.00	0.88 [0.83, 0.91]
Cuddliness	5.96 (0.59)	3.50–7.00	0.85 [0.81, 0.88]	5.51 (0.80)	1.00–6.86	0.87 [0.84, 0.89]	5.31 (0.81)	2.00–7.00	0.89 [0.86, 0.92]
Soothability	5.03 (0.70)	3.12–6.67	0.79 [0.74, 0.84]	5.13 (0.77)	2.94–6.72	0.86 [0.82, 0.89]	5.14 (0.85)	1.00–7.00	0.84 [0.78, 0.88]
Falling reactivity	5.19 (0.79)	2.50–6.90	0.78 [0.72, 0.82]	5.23 (0.87)	2.10–7.00	0.82 [0.77, 0.86]	5.30 (0.86)	2.00–7.00	0.80 [0.74, 0.84]

TABLE 3 | Fit indices of CFA models using 4-, 8-, and 12-month data.

Age	Model	*df*	χ^2	TLI	CFI	RMSEA	90% CI (RMSEA)
4 months	Three factor solution (Gartstein and Rothbart 2003)	74	323.85	0.71	0.762	0.113	0.10–0.13
	Modified three factor model	65	153.91	0.88	0.915	0.072	0.06–0.09
	Modified two factor model	69	160.96	0.88	0.912	0.071	0.06–0.09
8 months	Three factor solution (Gartstein and Rothbart 2003)	74	325.88	0.65	0.713	0.124	0.11–0.14
	Modified three factor model	66	149.74	0.87	0.904	0.076	0.06–0.09
	Modified two factor model	67	152.01	0.87	0.903	0.076	0.06–0.09
12 months	Three factor solution (Gartstein and Rothbart 2003)	74	317.01	0.58	0.659	0.127	0.11–0.14
	Modified three factor model	63	133.46	0.86	0.901	0.074	0.06–0.10
	Modified two factor model	64	135.47	0.86	0.901	0.073	0.06–0.09

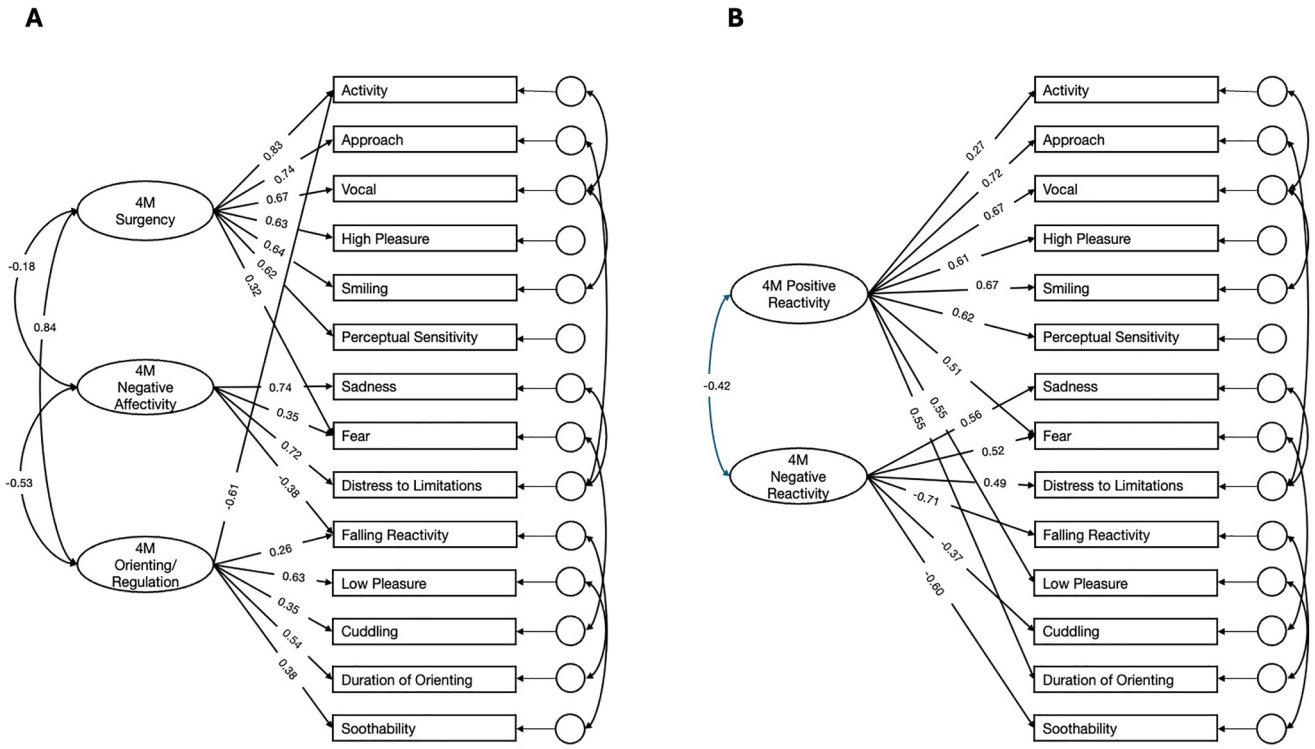

FIGURE 1 | Modified three-factor and two-factor models for 4-month data. 1A is the modified three-factor model; 1B is the modified two-factor model.

under the orienting/regulation factor to cross-load with both surgency and negative affect factors (i.e., positive and negative reactivity) as the covariance between orienting/regulation and negative affectivity was also significant. We then proceeded to remove loadings that were not significant to improve model fit. Figures 1–3 contain the modified three- and two-factor models at each timepoint. Based on fit indices, the two-factor models demonstrated similar fit to the modified three-factor models (see Table 3).

3.3 | Measurement Invariance by Demographic Characteristics

We examined measurement invariance of the modified three-factor model by infant sex (as assigned at birth), household income, and infant racial-ethnic majority/minority status using the modified three-factor solutions for each timepoint. For household income, we stratified income at $30,000, the federal poverty level for a family of 4. We tested the modified three-factor model instead of the exploratory two-factor model given the prevalent use of the three-factor model in the field. Below, we describe results of measurement invariance testing by demographic characteristics, and measurement models of configural or metric variance can be found in Supporting Information S1 (Figures S1–S7).

3.3.1 | Infant Sex

Results of measurement invariance testing by infant sex are presented in Table 4. Findings were not consistent across

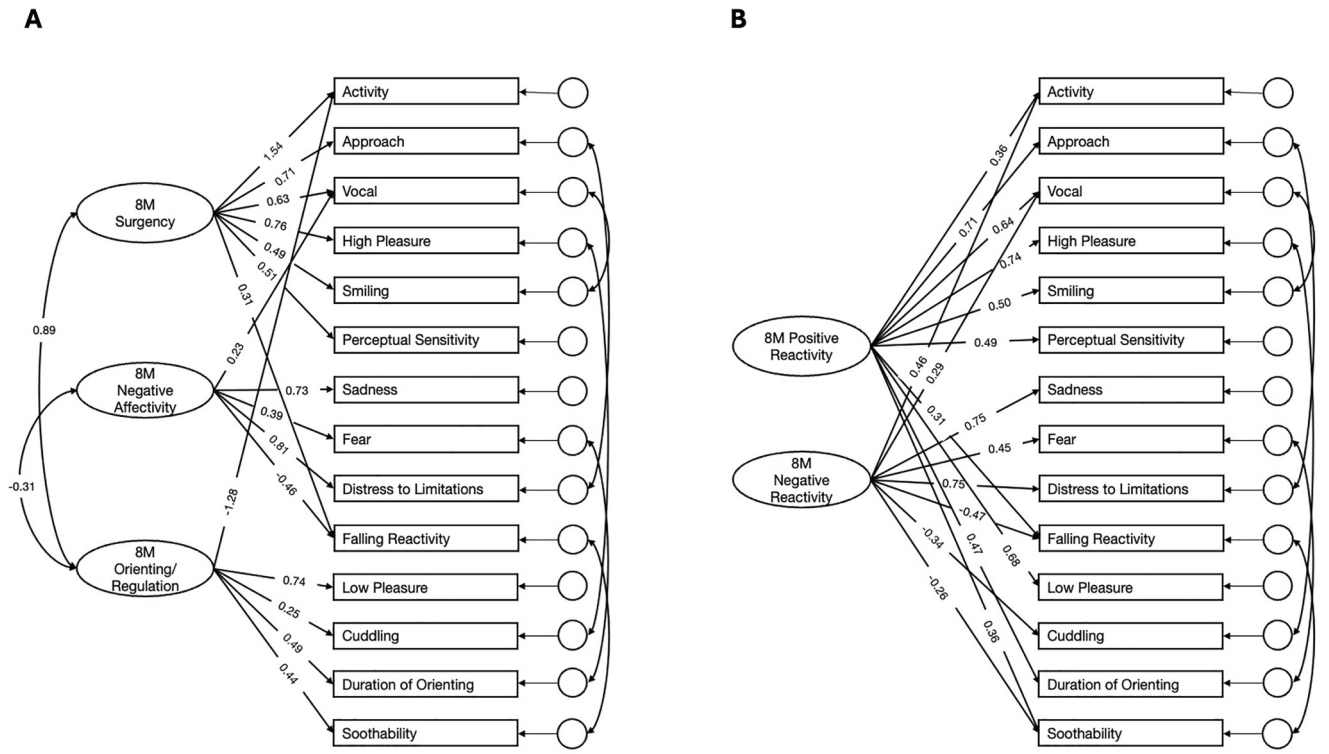

FIGURE 2 | Modified three-factor and two-factor models for 8-month data. 2A is the modified three-factor model; 2B is the modified two-factor model.

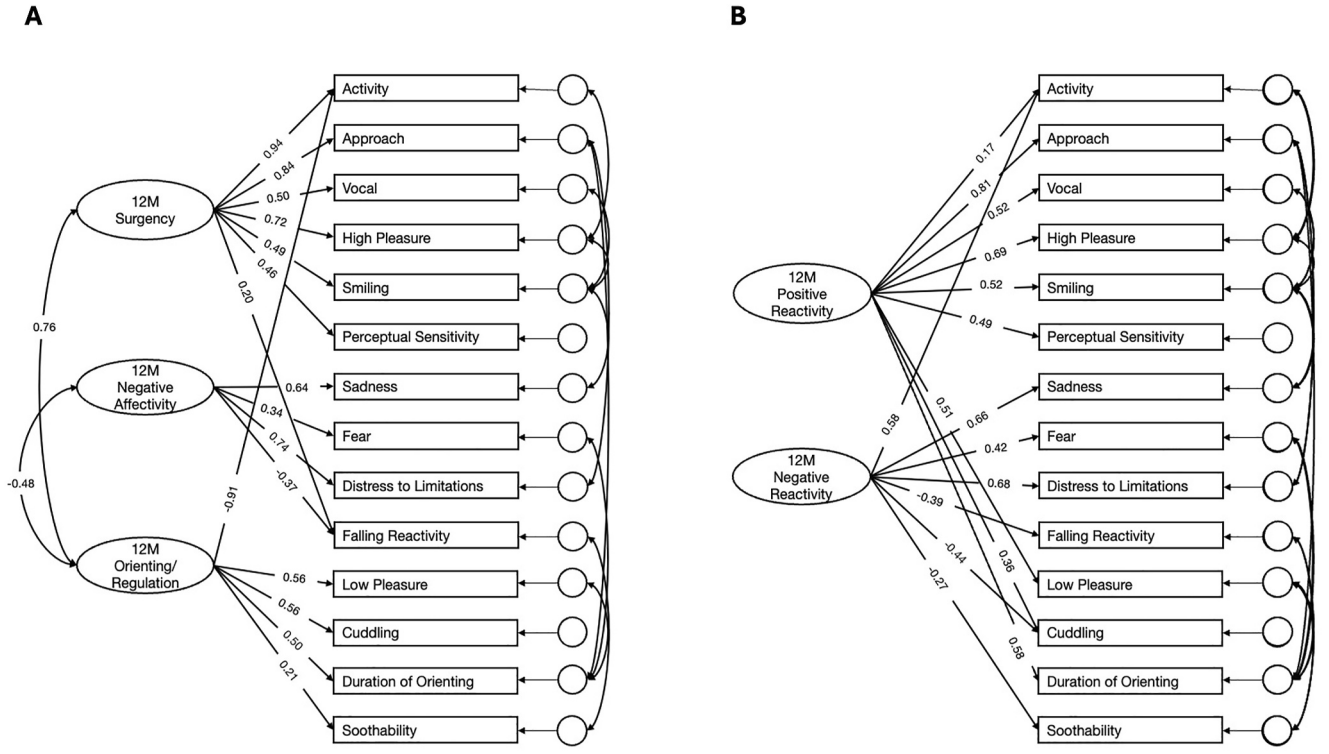

FIGURE 3 | Modified three-factor and two-factor models for 12-month data. 3A is the modified three-factor model; 3B is the modified two-factor model.

timepoints. At 4 months, the model fit across all groups was acceptable (χ^2 = 227.96, p < 0.001, CFI = 0.91, TLI = 0.87, RMSEA = 0.08), which means that the overall factor structure holds up similarly across infant sex. However, the p-value for the test comparing configural and metric invariance was significant; thus, equal factor loadings across sex are not

TABLE 4 | Results from measurement invariance testing of modified three-factor model (1B, 2B, and 3B) by infant sex.

Age	Model	df	CFI	RMSEA	AIC	BIC	χ^2	χ^2 diff.	df diff.	p
4 months	Configural	130	0.91	0.078	8247.1	8625.3	227.96			
	Metric invariance	144	0.89	0.080	8248.1	8578.1	257.76	29.80	14	0.008*
	Scalar invariance	155	0.88	0.080	8246.1	8536.7	276.89	19.12	11	0.059
8 months	Configural	132	0.92	0.072	6855.7	7210.0	204.11			
	Metric invariance	146	0.91	0.070	6845.3	7152.8	221.74	17.63	14	0.224
	Scalar invariance	157	0.89	0.076	6852.2	7122.9	250.62	28.88	11	0.002*
12 months	Configural	126	0.85	0.101	6062.5	6423.8	245.78			
	Metric invariance	139	0.83	0.101	6062.5	6381.8	271.72	25.93	13	0.017*
	Scalar invariance	150	0.82	0.102	6062.8	6346.7	294.08	22.37	11	0.022*

*$p < 0.05$.

TABLE 5 | Results from measurement invariance testing of modified three-factor model (1B, 2B, and 3B) by household income (stratified at $30,000, the federal poverty level for a family of 4).

Age	Model	df	CFI	RMSEA	AIC	BIC	χ^2	χ^2 diff.	df diff.	p
4 months	Configural	130	0.91	0.073	7880.1	8255.1	211.36			
	Metric invariance	144	0.92	0.066	7859.0	8185.4	218.30	6.94	14	0.937
	Scalar invariance	155	0.89	0.074	7874.3	8162.5	255.60	37.30	11	< 0.001*
8 months	Configural	132	0.92	0.078	6619.4	6971.6	198.88			
	Metric invariance	146	0.91	0.070	6610.9	6916.6	218.36	19.47	14	0.148
	Scalar invariance	157	0.87	0.082	6634.5	6903.7	264.05	45.69	11	< 0.001*
12 months	Configural	126	0.90	0.082	5608.0	5964.9	202.63			
	Metric invariance	139	0.87	0.088	5615.6	5930.1	235.25	32.62	13	0.002*
	Scalar invariance	150	0.82	0.100	5641.1	5921.6	283.74	48.49	11	< 0.001*

*$p < 0.05$.

supported. At 8 months, the *p*-value for the test comparing configural to metric invariance was not significant, and equal factor loadings across sex was supported. However, as the test comparing metric to scalar invariance is significant, strong invariance was not supported. We therefore cannot compare the values of the latent means across the two groups when using the current factor structures. Lastly, the model fit indices for a configural model across infant sex at 12 months were not acceptable ($\chi^2 = 245.76$, $p < 0.001$, CFI = 0.85, TLI = 0.78, RMSEA = 0.10), indicating that the overall factor structure does not hold up similarly across infant sex. Modification indices suggest that cuddliness should load onto negative affectivity for girls, but replication is needed in larger samples.

3.3.2 | Household Income

Table 5 contains the measurement invariance testing results by household income. At 4 and 8 months, equal factor loadings across family income were supported. However, as the test comparing metric to scalar invariance was significant, latent means across the two groups are not comparable. At 12 months, the fit indices for the configural model across household income was acceptable ($\chi^2 = 202.63$, $p < 0.001$, CFI = 0.90, TLI = 0.85, RMSEA = 0.08), which suggests that the overall factor structure holds up similarly across the groups, although equal factor loads were not supported.

3.3.3 | Infant Racial-Ethnic Majority/Minority Status

We also examined measurement invariance of the three-factor model by infant racial-ethnic majority/minority status (Table 6). At 4 and 8 months, equal factor loadings across groups were supported, though scalar invariance was not supported. However, at 12 months, the model fit indices for configural model showed that the factor structure did not hold up similarly across groups ($\chi^2 = 241.29$, $p < 0.001$, CFI = 0.86, TLI = 0.79, RMSEA = 0.099), suggesting that subscales may be loading onto different factors dependent on infant race/ethnicity. Modification indices suggest that smiling loads onto negative affectivity for infants who are White, non-Hispanic while soothability loads onto negative affect for infants from minority racial-ethnic groups, though replication is needed in larger samples.

4 | Discussion

The present study examined the replicability of the traditional, three-factor structure of temperament in the Infant Behavior Questionnaire using data from a longitudinal, multi-site study

TABLE 6 | Results from measurement invariance testing of modified three-factor model (1B, 2B, and 3B) by infant racial-ethnic majority/minority status.

Age	Model	*df*	CFI	RMSEA	AIC	BIC	χ^2	χ^2 diff.	*df* diff.	*p*
4 months	Configural	130	0.91	0.073	8126.2	8504.3	214.43			
	Metric invariance	144	0.91	0.068	8110.3	8439.4	226.49	12.06	14	0.601
	Scalar invariance	155	0.84	0.089	8166.2	8456.8	304.40	77.91	11	< 0.001*
8 months	Configural	132	0.92	0.069	6762.6	7116.9	196.76			
	Metric invariance	146	0.91	0.069	6757.4	7064.9	219.55	22.79	14	0.064
	Scalar invariance	157	0.87	0.083	6773.7	7044.5	269.66	50.10	11	< 0.001*
12 months	Configural	126	0.86	0.099	5913.5	6274.8	241.29			
	Metric invariance	139	0.81	0.107	5939.4	6258.8	293.20	51.91	13	< 0.001*
	Scalar invariance	150	0.73	0.125	5993.7	6277.6	369.46	76.26	11	< 0.001*

*$p < 0.05$.

in the United States. Our results are consistent with prior studies that find lack of replication for the traditional three-factor structure. Although we were able to achieve acceptable fit indices for three-factor models after allowing for cross-loadings, there was significant covariance between orienting/regulating and surgency factors, which led us to then explore modified two-factor models. Additionally, our findings highlight the need to examine measurement invariance in temperament measures such as the IBQ, underscoring the importance of robust measurement in studying early infant behaviors in studies inclusive of more diverse families.

We were unable to achieve a good fit with modified models unless we allowed for double loadings. While these cross-loadings make the factors harder to interpret, it may accurately reflect that temperamental traits may not always be orthogonal in behavior. Additionally, there are some key differences between the sample demographics that may explain our inability to replicate the traditional three-factor structure. For example, Gartstein and Rothbart (2003) had a much wider range of ages in determining the factors, while we aimed to replicate factor structure within single timepoints. This is important to consider given how manifestations of temperament may change across infancy due to developmental processes—for example, activity level increases throughout infancy (Buss and Plomin 1975).

In our study, fear positively loaded onto both surgency and negative affectivity in the modified models at 4 months. However, at 8 and 12 months, fear only loaded onto negative affectivity and was no longer loading onto the surgency factor. As fear develops, it may more strongly reflect individual differences in negative affectivity as opposed to reactivity more broadly. Additionally, Gartstein et al. (2006) examined differences in stability in IBQ subscales in 3-, 6-, and 9-month-old infants from the United States, Spain, and China. Findings showed that there were differences in stability across certain subscales (e.g., distress to limitations, duration of orienting) across these three groups, highlighting that culture may play a role in developmental change in temperament traits as well. We were limited by power to conduct longitudinal invariance testing; however, our study provides preliminary evidence that there may be changes in factor structure across development as the subscales cross-loaded differently on temperament factors at each timepoint. Dias et al. (2021) found evidence for stability in the IBQ factor structure at 2 weeks, 3, 6, and 12 months in a Portuguese sample; however, it should be noted that their sample demographics differed significantly from ours, and they did not examine cross-loadings across multiple timepoints. Future work should examine change in factor structure across development across diverse populations to better understand how different dimensions of reactivity and regulation may develop across infancy. Our results should also be interpreted with the caveat that maternal education was significantly associated with attrition at 8 and 12 months. While there are developmental changes reflected in our data, it may also be the case that it is driven by demographic differences due to attrition.

The modified three-factor models also showed that the broader three factors were not orthogonal. More specifically, orienting/regulation was associated with both surgency and negative affectivity. This may highlight the interplay and reciprocal associations between reactive and regulatory processes during this period of development. While they may be conceptually distinct, it should be noted that the covariance between surgency and regulation was very high across all three ages in our sample. The high covariance may indicate lack of discriminant validity between surgency and orienting factors. To address this issue, we examined a modified two-factor model of positive and negative reactivity that includes subscales from the regulatory factor. It could be that it is difficult at these ages for parents to distinguish between reactive and regulatory behaviors, especially as infants are still developing their regulatory capabilities. As we did not have a priori hypotheses about a two-factor model, this was an exploratory alternate structure which will require empirical replication in other samples.

Sample demographics are also important to consider when examining dimensions of infant temperament, especially given the need for more inclusive research, and including increased representation of families from marginalized groups. Although we were able to identify modified factor structures with acceptable fit across all timepoints, it was evident there is no measurement invariance across sociodemographic characteristics. More specifically, there was generally support for metric invariance (i.e., subscales contributed to factors similarly across groups) at 4 and 8 months. However, the factor

structure did not replicate across infant sex and infant race/ethnicity at 12 months. Enlow et al. (2016) found that factor means differed across White, Hispanic, and Black/Haitian participants in their study. However, in our study, we found that more fine-grained temperament traits (as assessed by the subscales on the IBQ) may contribute differently to the latent constructs of surgency, negative affectivity and orienting depending on infants' sex and racial-ethnic majority/minority status at 12 months. This lack of invariance at 12 months may reflect differences in postnatal experiences across groups that may modulate developmental changes in temperament, though it may also be due to differences in our participants due to attrition. Future studies should consider how sociocultural contextual factors may contribute to differences in the structure of temperament across infancy.

Due to limitations in our sample size, we could not focus on specific racial/ethnic groups or take a within-group approach to better understand the differences in the overarching temperament factors in racial-ethnic minority groups. For example, it is likely there may be some confounding of race/ethnicity and language in our sample. Additionally, it is important to note that we used household income, which is an imperfect measure of economic stress and hardship compared to other metrics such as income-to-needs ratios. However, we hope that the present study will call attention to how sociodemographic variables, such as race/ethnicity and socioeconomic status, are important to consider in measurement issues around temperament. Future studies should aim to examine how contextual processes within and across racial-ethnic minority groups may be associated with infant temperament.

One limitation of the current study is that we evaluated measurement invariance by testing the factor structure by bivariate groups without consideration of other correlates or moderators. This is an important limitation to consider, as it does not account for the intersectional identities of families. Future work with larger sample sizes could utilize methods such as moderated nonlinear factor analysis (Bauer 2017), which would allow for simultaneous evaluation of measurement invariance over multiple background variables.

These results have important implications for future research. First and foremost, researchers using parent-report measures of temperament such as the IBQ should examine the factor structure within their datasets for the three broader temperament factors of surgency, negative affectivity, and orienting. Without this step, researchers are making the assumption that (1) their data fit the three-factor structure well, and (2) the constructs of surgency, negative affectivity, and orienting are invariant (i.e., equal) across different groups or measurement occasions. Without testing these assumptions, researchers may be adding noise and can contribute to errors in conclusions regarding how temperament may be associated with later outcomes. These assumptions can be problematic, especially in larger, multi-site studies collected from culturally and sociodemographically diverse families. As developmental research moves toward greater inclusivity, it is imperative to interrogate the measurement tools we use to assess infant behaviors and ensure that we are accurately capturing temperament constructs. If the factor structure does not replicate, it may be better to create composites based on conceptual, theoretical justification in line with research questions. For example, we used the subscales of sadness, distress to limitations, and fear to create a composite of negative emotionality given our interests in associations with maternal internalizing symptoms (Zhou et al. 2023).

An alternative method for characterizing infant temperament could rely on person-centered analyses to identify different temperament profiles (e.g., Beekman et al. 2015). For example, Beekman et al. (2015) identified different temperament profiles using the IBQ subscales of activity level, distress to novelty, distress to limitations, duration of orienting, soothability, and smiling and laughing. This data-driven approach would allow researchers to identify latent sub-groups of temperament profiles without assuming that specific subscales contribute to broader factor scales.

In conclusion, our study examined the replicability of the traditional, three-factor structure of temperament in the Infant Behavior Questionnaire using data from a longitudinal, multi-site study in the United States. Our results are consistent with prior work that the widely used factor structure does not always fit the data well. Our findings highlight the importance of considering measurement invariance in temperament measures, and that it may not be appropriate to compare factor means when examining temperament in a diverse sample. Researchers should examine factor structure thoughtfully in their own datasets before creating composites and consider how sociodemographic characteristics of their samples may impact their data.

Author Contributions

Anna M. Zhou: conceptualization, formal analysis, methodology, visualization, writing – original draft, writing – review and editing. **Koraly Pérez-Edgar:** funding acquisition, writing – review and editing. **Vanessa LoBue:** funding acquisition, writing – review and editing. **Kristin A. Buss:** conceptualization, funding acquisition, writing – review and editing.

Acknowledgments

Data for this study is supported by grants from National Institutes of Health awarded to Pérez-Edgar, Buss, and LoBue (R01MH109692). Dr. Zhou was supported by a postdoctoral training grant (T32 MH015442). Drs. Buss and Pérez-Edgar's Psychology Professorships are supported by the Social Science Research Institute of The Pennsylvania State University, and endowments through the Tracy Winfree and Ted H. McCourtney Professorship in Children, Work, and Families (Buss) and the McCourtney Professorship of Child Studies (Pérez-Edgar). We would like to thank the families involved with the study, as well as the staff and larger research team for their contributions to the study. We would especially like to thank and acknowledge Dr. Brendan Ostlund for his contributions.

Ethics Statement

The Institutional Review Boards at the Pennsylvania State University (IRB STUDY00004097) and Rutgers University (#Pro-2020000418) approved all procedures. Parents provided written consent and were compensated for their participation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are publicly accessible on Databrary at http://doi.org/10.17910/b7.1288 (LoBue et al. 2021).

References

Bauer, D. J. 2017. "A More General Model for Testing Measurement Invariance and Differential Item Functioning." *Psychological Methods* 22, no. 3: 507–526. https://doi.org/10.1037/met0000077.

Beekman, C., J. M. Neiderhiser, K. A. Buss, et al. 2015. "The Development of Early Profiles of Temperament: Characterization, Continuity, and Etiology." *Child Development* 86, no. 6: 1794–1811. https://doi.org/10.1111/cdev.12417.

Buss, A. H., and R. Plomin. 1975. *A Temperament Theory of Personality Development*. Wiley-Interscience.

Dias, C. C., R. Costa, T. M. Pinto, and B. Figueiredo. 2021. "The Infant Behavior Questionnaire – Revised: Psychometric Properties at 2 Weeks, 3, 6 and 12 Months of Life." *Early Human Development* 153: 105290. https://doi.org/10.1016/j.earlhumdev.2020.105290.

Enlow, M. B., M. T. White, K. Hails, I. Cabrera, and R. J. Wright. 2016. "The Infant Behavior Questionnaire-Revised: Factor Structure in a Culturally and Sociodemographically Diverse Sample in the United States." *Infant Behavior and Development* 43: 24–35. https://doi.org/10.1016/j.infbeh.2016.04.001.

Gartstein, M. A., C. Gonzalez, J. A. Carranza, et al. 2006. "Studying Cross-Cultural Differences in the Development of Infant Temperament: People's Republic of China, the United States of America, and Spain." *Child Psychiatry and Human Development* 37, no. 2: 145–161. https://doi.org/10.1007/s10578-006-0025-6.

Gartstein, M. A., G. G. Knyazev, and H. R. Slobodskaya. 2005. "Cross-Cultural Differences in the Structure of Infant Temperament: United States of America (US) and Russia." *Infant Behavior and Development* 28, no. 1: 54–61. https://doi.org/10.1016/j.infbeh.2004.09.003.

Gartstein, M. A., and J. Marmion. 2008. "Fear and Positive Affectivity in Infancy: Convergence/Discrepancy Between Parent-Report and Laboratory-Based Indicators." *Infant Behavior and Development* 31, no. 2: 227–238. https://doi.org/10.1016/j.infbeh.2007.10.012.

Gartstein, M. A., and M. K. Rothbart. 2003. "Studying Infant Temperament via the Revised Infant Behavior Questionnaire." *Infant Behavior and Development* 26, no. 1: 64–86. https://doi.org/10.1016/s0163-6383(02)00169-8.

Hu, L. T., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling* 6, no. 1: 1–55. https://doi.org/10.1080/10705519909540118.

Kass, R. E., and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90, no. 430: 773–795. https://doi.org/10.2307/2291091.

LoBue, V., K. Perez-Edgar, and K. Buss. 2021. "Publications From the Longitudinal Attention and Temperament Study (LANTS)." *Databrary*. https://www.databrary.org/volume/1288.

Moshagen, M., and M. Bader. 2024. "semPower: General Power Analysis for Structural Equation Models." *Behavior Research Methods* 56, no. 4: 2901–2922. https://doi.org/10.3758/s13428-023-02254-7.

Ostlund, B., S. Myruski, K. Buss, and K. E. Pérez-Edgar. 2021. "The Centrality of Temperament to the Research Domain Criteria (RDoC): The Earliest Building Blocks of Psychopathology." *Development and Psychopathology* 33, no. 5: 1584–1598. https://doi.org/10.1017/s0954579421000511.

Parade, S. H., and E. M. Leerkes. 2008. "The Reliability and Validity of the Infant Behavior Questionnaire-Revised." *Infant Behavior and Development* 31, no. 4: 637–646. https://doi.org/10.1016/j.infbeh.2008.07.009.

Pérez-Edgar, K., V. LoBue, K. A. Buss, A. P. Field, and LAnTs Team. 2021. "Study Protocol: Longitudinal Attention and Temperament Study." *Frontiers in Psychiatry* 12: 656958. https://doi.org/10.3389/fpsyt.2021.656958.

Rosseel, Y. 2012. "Lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48, no. 2: 1–36. https://doi.org/10.18637/jss.v048.i02.

Rothbart, M. K. 1981. "Measurement of Temperament in Infancy." *Child Development* 52, no. 2: 569–578. https://doi.org/10.2307/1129176.

Rothbart, M. K. 1989. "Temperament in Childhood: A Framework." In *Temperament in Childhood*, edited by G. A. Kohnstamm, J. E. Bates, and M. K. Rothbart, 59–73. John Wiley & Sons.

Rothbart, M. K., and J. E. Bates. 2006. "Temperament." In *Handbook of Child Psychology: Volume 3 Social, Emotional, and Personality Development*, edited by N. Eisenberg, W. Damon, and R. M. Lerner, 99–166. John Wiley & Sons.

Shiner, R. L., K. A. Buss, S. G. McClowry, S. P. Putnam, K. J. Saudino, and M. Zentner. 2012. "What Is Temperament now? Assessing Progress in Temperament Research on the Twenty-Fifth Anniversary of Goldsmith et al." *Child Development Perspectives* 6, no. 4: 436–444. https://doi.org/10.1111/j.1750-8606.2012.00254.x.

Zhou, A. M., M. N. Lytle, E. A. Youatt, K. Pérez-Edgar, V. LoBue, and K. A. Buss. 2023. "Examining Transactional Associations Between Maternal Internalizing Symptoms, Infant Negative Emotionality, and Infant Respiratory Sinus Arrhythmia." *Biological Psychology* 182: 108625. https://doi.org/10.1016/j.biopsycho.2023.108625.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.