Establishing the Measurement Invariance of the Very Short Form of the Infant Behavior Questionnaire Revised for Mothers Who Vary on Race and Poverty Status

By: Esther M. Leerkes, Jinni Su, Beth A. Reboussin, Stephanie S. Daniel, Chris C. Payne, and Joseph G. Grzywacz

Leerkes, E.M., Su, J. Reboussin, B.A., Daniel, S.S., Payne, C.C., & Grzywacz, J.G. (2017) Establishing the measurement invariance of the Very Short Form of the Infant Behavior Questionnaire Revised for mothers who vary on race and poverty status. *Journal of Personality Assessment, 99,* 94-103. https://doi.org/10.1080/00223891.2016.1185612

Abstract:

We examined the measurement invariance of the Infant Behavior Questionnaire Revised–Very Short Form (IBQR–VSF; Putnam, Helbig, Gartstein, Rothbart, & Leerkes, 2014) in a sample of 470 racially (185 White, 285 African American) and socioeconomically diverse mothers (158 below federal poverty threshold, 296 above federal poverty threshold) of infants. Using multigroup confirmatory factor analysis, we demonstrated configural, full metric, and full scalar invariance demonstrating that the 3-factor structure (negative emotionality, positive affectivity/surgency, orienting/regulatory capacity), pattern of item loadings, and item means were comparable for White and African American mothers, and for poor and not poor mothers. In addition, we demonstrated full error invariance across racial groups and partial error variance invariance across poverty status, demonstrating that item reliability was comparable for White and African American mothers, and both those above and below the poverty line (with the exception of a subset of items). Thus, the IBQR–VSF appears appropriate for use in racially and socioeconomically diverse samples.

Keywords: Infant Behavior Questionnaire | IBQR–VSF | parenting | infant temperament

Article:

Infant temperament is defined as biologically based individual differences in infants' reactivity and self-regulation in response to environmental stimuli (Rothbart & Bates, 2006). The reactivity component includes the latency, intensity, and duration of affective, attentional, and motor responses to stimuli. The regulatory component includes the manner and ability with which these reactions are modulated (Rothbart & Bates, 2006). Infant temperament has been widely studied since the 1970s (e.g., Carey & McDevitt, 1978; Thomas & Chess, 1977), and early individual

differences in temperament are correlated with children's later social and emotional outcomes, family functioning, parental well-being, and the quality of parenting (Rothbart & Bates, 2006). Parental reports of temperament are commonly used given their ease of implementation, cost-effectiveness, and capitalization on parents' opportunity to observe their infants over time and across settings. One such measure is the Infant Behavior Questionnaire Revised–Very Short Form (IBQR–VSF; Putnam, Helbig, Gartstein, Rothbart, & Leerkes, 2014), which is gaining increased use because of its brevity and focus on both temperamental reactivity and regulation. Prior research with the IBQR–VSF has been conducted in primarily White samples (Putnam et al., 2014). As such, the appropriateness of this measure for diverse samples remains unclear. We examine the extent to which the IBQR–VSF is invariant between White and African American mothers and mothers above and below federal poverty guidelines.

Infant temperament as a central construct in developmental science and family studies

Infant temperamental reactivity (both negative and positive emotionality) and regulation are consistent correlates of a host of child, parent, and family outcomes. Much of this research has focused on the negative emotionality component of reactivity because it is conceived as a risk factor for infants and their families. For example, infants and toddlers rated as fussy or prone to frustration, based in part on maternal reports on the Infant Behavior Questionnaire (IBQ; Rothbart, 1981) demonstrate concurrent deficits in self-regulation via lower self-distraction (r = –.30; Calkins, Dedmon, Gill, Lomax, & Johnson, 2002) and are at greater risk for heightened conduct problems from age 3 to age 14 (r = .21; Lahey et al., 2008). Similarly, parental reports of negative emotionality obtained when children were 3 to 12 months of age using the Infant Behavior Questionnaire–Revised (IBQ–R) were correlated with both internalizing and externalizing symptoms when children were 2 and 4 years old (rs ranged from .17–.35; Gartstein, Putnam, & Rothbart, 2012).

In terms of the broader family, parental reports of negative emotionality on the IBQ or IBQ–R are linked with parents' elevated depressive symptoms (Solmeyer & Feinberg, 2011) and parenting stress (Oddi, Murdock, Vadnais, Bridgett & Gartstein, 2013) and with lower parenting efficacy (Leerkes & Burney, 2007; Solmeyer & Feinberg, 2011), parenting satisfaction (Leve, Scaramella, & Fagot, 2001; Mehall, Spinrad, Eisenberg, & Gaertner, 2009), marital satisfaction (Leve et al., 2001), adaptive coparenting (Burney & Leerkes, 2010), and sensitive parenting behavior (Bridgett et al., 2009; Calkins, Hungerford, & Dedmon, 2004). These main effect associations range from r = .20 to .40 with a mean of .27, absolute value. Moreover, links between infant negative emotionality and negative family outcomes are particularly apparent when other risk factors are present in the family system as evidenced by modest but statistically significant interaction effects (Burney & Leerkes, 2010; Crockenberg & Leerkes, 2003; Schoppe-Sullivan, Mangelsdorf, Brown, & Sokolowski, 2007). Importantly, links between infant temperament and family functioning are likely bidirectional (e.g., Bridgett et al., 2009).

More recently researchers have used the IBQ or IBQ–R to focus on the regulatory component of temperament as well as positive emotionality and activity level in relation to child outcomes and family functioning. For example, parents' reports of infant regulation are linked with children's better effortful control (i.e., the ability to control attention and behavior) over time (Bridgett et al., 2011) and with higher coparenting quality (Burney & Leerkes, 2010; Solmeyer &

Feinberg, 2011) and parenting self-efficacy (Leerkes & Burney, 2007; Solmeyer & Feinberg, 2011). In addition, parent-reported infant soothability buffers observed maternal sensitivity (assessed during a 1-hr home visit) and maternal self-efficacy from the negative effects of infant negative emotionality such that the negative association between infant negative emotionality and sensitivity or efficacy is significant when infant regulation is low but not when infant regulation is high (Crockenberg & Leerkes, 2003; Ghera, Hane, Malesa, & Fox, 2006). These results suggest that early parent-reported regulation is linked with optimal infant adjustment and family functioning. In contrast, the results are mixed when it comes to parent-reported infant positive affect or surgency. Positive affect or surgency has been linked with declining observed maternal sensitivity and increasing intrusiveness over time (rated during the brief engage and reengage episodes of the still face procedure at three time points; Planalp, Braungart-Rieker, Lickenbrock & Zentall, 2013), but also with children's lower risk of conduct problems (Lahey et al., 2008) and with better effortful control (Gartstein, Slobodskaya, Putnam, & Kinsht, 2009) over time. These associations ranged from r = .10 to .33 (mean r = .24). Most of this research has been conducted in primarily middle-income, White samples, and in more diverse samples, race and income were covaried (e.g., Calkins et al., 2002; Lahey et al., 2008).

History and development of the IBQR–VSF

Given the frequency and general consistency with which parental perceptions of infant temperament are related with important child, parent, and family outcomes, ensuring that parent perceptions of infant temperament are adequately measured is essential. With this goal in mind, Rothbart (1981) created the first version of the IBQ, which included six scales: Activity Level, Smiling and Laughter, Fear, Distress to Limitations, Duration of Orienting, and Soothability. Later, the Vocal Reactivity scale was added (Rothbart, 1986). Parents report only on infant behavior in the recent past and rate the frequency of specific behaviors in well-defined contexts (e.g., when introduced to a stranger, riding in the car, waiting in the crib). This measure was frequently used in developmental research. Then Gartstein and Rothbart (2003) published the IBQ–R. The IBQ–R incorporated new items in an effort to capture additional specific dimensions of temperament apparent in studies of older children and adults including approach, high pleasure, perceptual sensitivity, sadness, low pleasure, falling reactivity, and cuddliness, resulting in a longer instrument (191 items) to tap these 14 distinct dimensions. Factor analyses of the 14 scale scores suggested three overarching broadband dimensions of temperament: negative affectivity (sadness, distress to limitations, fear, and falling reactivity reversed), surgency/extraversion (approach, vocal reactivity, high-intensity pleasure, smiling and laughter, activity level, perceptual sensitivity), and orienting/regulation (low-intensity pleasure, cuddliness, duration of orienting, and soothability).

Recently, Putnam et al. (2014) noted the advantages of a fine-grained approach to assessing temperament, but also acknowledged that such a long measure is problematic for researchers for whom temperament is not the central focus. Thus, they created the 89-item IBQR–Short Form (SF) and the 37-item IBQR–VSF. A key difference between the two, in addition to length, is that the IBQR–SF assesses all 14 temperament dimensions, whereas the IBQR–VSF was designed to assess only the three broadband dimensions. The three broadband scales maintained the original structure described by Gartstein and Rothbart (2003) but were renamed positive affectivity/surgency, negative emotionality, and orienting/regulatory capacity. Drawing on data

from 11 independent samples, Putnam et al. (2014) demonstrated the IBQR–VSF had adequate internal reliability (Cronbach's alpha ranged from .75–.78), stability over 2 months (*r*s ranged from .52–.59), and interparental agreement (*r*s ranged from .28–.61). Based on the psychometric characteristics of the IBQR–VSF and its brevity, the authors concluded that it is ideal for large-scale and epidemiological studies.

Rationale for this study

To date, the extent to which the factor structure and item loadings of the IBQR–VSF are invariant across racial and income groups has not been demonstrated. Although Putnam et al. (2014) reported similar and more than adequate internal consistency reliability for each subscale of the IBQR–VSF for separate racial and ethnic groups, it is possible that more subtle, but meaningful, differences in the item loadings or means that could affect interpretation of results are apparent between groups. Establishing measurement invariance of the IBQR–VSF is particularly important for studies with diverse samples that involve comparisons between groups because lack of invariance could lead to biased interpretations of group differences in construct means or associations. That is, if the IBQR–VSF is not invariant across racial or income groups, infant temperament as measured by the IBQR–VSF might have different meanings across groups; as a result, any findings regarding group differences in infant temperament as well as its relationship with other constructs might be over- or underestimated due to measurement bias.

In four studies that used the IBQ or IBQ–R in diverse samples, minority (primarily African American) parents and low-income parents rated their infants higher on negative emotionality dimensions than did White and higher income parents (Calkins et al., 2002; Chen, 2012; Garrett-Peters, Mills-Koonce, Adkins, Vernon-Feagans, & Cox, 2008; Parade & Leerkes, 2008). Across studies, these differences were small to moderate in magnitude (*r*s ranged from .15–.35). Interpreting these differences is difficult without knowledge of measurement invariance: They could reflect meaningful differences in infant behavior, or they could reflect underlying differences in how parents rate these items based on race and income. African American and White adults hold different beliefs about the expression and control of emotions, the role of emotions in daily life, and appraisals of emotion intensity (Cole & Tan, 2007; Matsumoto, 1993; Parker et al., 2012), all of which could contribute to differences in the perception and reporting of infant temperament. For example, African Americans tend to rate facial expressions of distress as more intense (Matsumoto, 1993) and report it is less socially acceptable to express negative emotions (Matsumoto, 1989; Nelson, Leerkes, O'Brien, Calkins, & Marcovitch, 2012) than other groups. In regard to income, lower income mothers might interpret items differently or might have fewer opportunities to observe their infant in certain settings relative to higher income mothers (Putnam & Rothbart, 2006). There is also some evidence that lower income mothers have more negative beliefs about infant crying ($r = .13$; Leerkes et al., 2015) and are somewhat less skilled at accurately perceiving specific infant negative emotions ($r = –.23$ and $–.17$, respectively; Bernstein, Tenedios, Laurent, Measelle, & Ablow, 2014; Leerkes et al., 2015) relative to higher income mothers, which might alter how mothers respond to items on the IBQR–VSF. Given evidence that emotion beliefs are linked with attention to emotion cues (Dennis & Halberstadt, 2013), African American mothers and low-income mothers might overrate the intensity and frequency of some infant distress cues because infant distress is counter to their preference for limited emotion expression, or they might overattend to and hence

overrate infant positive affect in some contexts because it is desirable. Thus, possible measurement invariance based on both race and poverty status is a concern.

However, prior research testing measurement invariance of personality or emotion-relevant reports of child behavior between racial and ethnic groups and poverty status groups have primarily demonstrated invariance. For example, teacher reports of preschooler's self-regulatory behavior (McCoy, Raver, Lowenstein, & Tirado-Strayer, 2011) and youth self-reports of behavioral and emotional functioning (Harrell-Williams, Raines, Kamphaus, & Dever, 2015) were demonstrated to be invariant across racial and income groups. In other studies, partial invariance has been demonstrated across racial groups on youth self-reports of anxiety and depression (Holly, Little, Pina, & Caterino, 2015; Trent et al., 2013). To our knowledge, no prior studies have addressed measurement equivalence of the IBQR–VSF or other parent-report measures of infant temperament. However, the measurement invariance of observational measures of fear and anger reactivity and regulation across race (White and African American) and income (poor and not poor) has been established (Willoughby, Stifter, & Gottfredson, 2015).

This study

In this study, we capitalize on data from two diverse samples of mothers who completed the IBQR–VSF when infants were 3 to 12 months old to evaluate measurement invariance of the IBQR–VSF for White and African American mothers and for poor and nonpoor mothers. We evaluate four increasingly restrictive levels of measurement invariance for the IBQR–VSF across racial groups and then across income groups: configural invariance, metric invariance, scalar invariance, and error variance invariance. These aspects of invariance are most commonly evaluated in the literature, and they each provide insight regarding whether and how a measure functions similarly or differently across groups (Byrne & Watkins, 2003; Cheung & Rensvold, 2002; Steenkamp & Baumgartner, 1998; Van de Schoot, Lugtig, & Hox, 2012). Configural invariance, sometimes called pattern invariance, refers to similarity in factor structure of the measure across groups (i.e., similar latent factors are represented across the groups). That is, the items comprising the measure exhibit the same configuration of salient (nonzero) and nonsalient (zero or near zero) factor loadings across different groups. Metric invariance, sometimes also referred to as weak factorial invariance, refers to equal strengths of the relations between scale items and their respective underlying construct across groups. That is, metric invariance indicates equal factor loadings across groups and implies that the same underlying constructs (i.e., latent factors) are being measured across groups. Scalar (intercept) invariance, also referred to as strong factorial invariance, indicates that individuals who have the same scores on the latent variables would have the same scores on the observed items, across groups. Scalar invariance implies that across-group differences in the means of the observed scale items are due to differences in the means of the underlying constructs represented by latent variables. Scalar invariance is required to compare factor means across groups. Error variance invariance, also referred to as strict factorial invariance, is defined as the same amount of measurement error across groups and implies that the scale items are equally reliable across groups.

Method

Participants

The sample for this project was drawn from two studies conducted in the same geographic region and time frame. The Triad Child Study is a prospective longitudinal study designed to identify predictors of maternal sensitivity and early infant well-being. The original sample consisted of 259 women expecting their first child. When infants were around 6 months old, the second observation and data collection point in this study, 226 participating mothers completed the IBQR–VSF. This observation occurred between June 2010 and February 2012. The Women, Work and Wee Ones Project is a prospective longitudinal study designed to examine the effects of maternal employment schedules on maternal well-being, parenting behavior, and infant adjustment in low-income families. The sample consisted of 285 mothers of 3-month-old children. When infants were 12 months old, 243 participating mothers completed the IBQR–VSF. These waves of data collection occurred between September 2010 and October 2013.

For mothers who did not complete the IBQR–VSF at the 12-month assessment or whose infants were 13 months or older at the completion of the IBQR–VSF in the second study (n = 121), data for the 3-month assessment were used. For mothers who completed the IBQR–VSF at the 3-month assessment and at the 12-month assessment when their infants were younger than 13 months old (n = 164), we randomly selected half of these mothers to use the 3-month data and the other half to use the 12-month data. We did the random draws three times, and thus created three analytic samples. We conducted parallel analysis three times, and the patterns of results were the same across the three analytic samples. We present results from analyses with one of the samples. The final analytic sample for this article consists of 470 (out of 511) mothers drawn from these two studies; 41 were excluded from the analytic sample because they were not White or African American. Thus, IBQR–VSF data were included from 176 three-month-old infants, 214 six-month-old infants, and 80 twelve-month-old infants.

The demographic characteristics of the complete analytic sample and each subsample are provided in Table 1. Most demographics varied between the two studies as a function of varying recruitment goals. For example, the Triad Child Study deliberately recruited an equal number of African American and White mothers; and the Women, Work and Wee Ones Study deliberately recruited low-income mothers. Importantly, the combined sample reflects a diverse group of mothers with a sufficient number of African American and White mothers and mothers living below and above the federal poverty level to facilitate the planned analyses.

Table 1. Sample characteristics.

	Combined samples[a]		**Triad Child Study[b]**		**Women, Work & Wee Ones[c]**		**Comparison**
	n	**%**	***n***	**%**	***n***	**%**	$\chi^2(df)$
Maternal race							$23.86(1)^{**}$
White	185	39.4	110	51.4	75	29.3	
African American	285	60.6	104	48.6	181	70.7	
Income to needs							$19.53(1)^{**}$
Below federal poverty	158	33.6	47	23.6	111	43.4	
Above federal poverty	296	63.0	152	76.4	144	56.3	
Not reported	16	3.4	15	7.0	1	0.4	
Maternal education							$60.31(2)^{**}$
HS/GED or less	132	28.1	50	23.4	82	32.0	
Some college/trade	206	43.8	66	30.8	140	54.7	

4-year degree or more	130	27.7	96	44.9	34	13.3	
Not reported	2	0.4	1	0.9	0	0	
Marital status							36.47(2)**
Married	141	30.0	92	43.0	49	19.1	
Marriage-like	66	14.0	32	15.0	34	13.3	
Other	262	55.7	89	41.6	173	67.6	
Not reported	1	0.2	1	0.5	0	0	
Infant gender							2.01(1)
Males	243	51.7	103	48.1	140	54.7	
Females	227	48.3	111	51.9	116	45.3	
Maternal age (years)[d]	26.56	5.33	25.98	5.44	27.05	5.19	2.19 (468)*
Infant age (months)[d]	6.22	3.28	6.40	.76	6.08	4.31	−1.09(468)

Note. HS = high school; GED = general equivalency diploma.
[a] $N = 470$. [b] $N = 214$. [c] $N = 256$. [d] Mean, standard deviation, and t value reported rather than N, %, and χ^2.
* $p < .05$. ** $p < .01$.

Procedure

In the Triad Child Study, expectant mothers were recruited at childbirth classes offered in the local hospital and public health department; breastfeeding classes offered through the Special Supplemental Nutrition Program for Women, Infants and Children (WIC); obstetric practices; and word of mouth. On enrollment in the study, women were mailed consent forms and questionnaires and completed an interview. Mothers and infants visited the laboratory for a videotaped observation of mother–infant interaction when infants were 6 months old and about 1 year old. Prior to the 6-month visit, mothers were mailed a packet of questionnaires including the IBQR–VSF and an updated demographics form that they returned at the visit.

In the Women, Work and Wee Ones Study, mothers were initially contacted by study staff at the maternity wards of three regional hospitals shortly after giving birth or in the waiting room of local WIC clinics. On enrollment in the study, a 3-month home visit was scheduled. During this visit, informed consent was obtained and mothers completed an interviewer-administered survey questionnaire and some self-administered instruments. When infants were 1 year old, a similar home visit was scheduled and interviewers administered the IBQR–VSF and an updated demographic form at this time. Procedures for both studies were approved by the institutional review board of the home institution.

Measures

Participants in both studies provided demographic information at each wave of data collection, including information about age, race, education level, income, household composition, marital status, employment status, and their infant's gender and health. When their infants were 3, 6, or 12 months old, participants in both studies completed the IBQR–VSF (Putnam et al., 2014). The IBQR–VSF is made up of three broadband scales including Surgency (13 items; e.g., How often during the week did your baby move quickly toward new objects?), Negative Affect (12 items; e.g., When tired, how often did your baby show distress?), and Effortful Control (12 items; e.g., Play with one toy for 5 to 10 minutes.). For each item, mothers are asked to rate the extent to which their child engaged in the target behavior during the past 7 days on a scale from 1 (*never*)

to 7 (*always*). In the event certain situations did not arise in the prior week (e.g., be introduced to an unfamiliar adult), *not applicable* is a response option. In prior research, the IBQR–VSF has demonstrated adequate internal consistency, test–retest reliability, and interrater agreement between mothers and fathers (Putnam et al., 2014).

An income-to-needs ratio was calculated for each participant by dividing their total family income by the federal poverty threshold for a family of that size. Scores below 1 indicate that a family's income is below the federal poverty level. Scores above 1 indicate that the family's income is above the federal poverty level. We classified participants into low-income (income-to-needs ratio below 1) and high-income (income-to-needs ratio above 1) groups. Poverty and minority status correlated modestly in the combined sample, $r = .21$, $p < .05$.

Analysis

A series of multigroup confirmatory factor analyses (MGCFAs) using M*plus* version 7 was used to evaluate measurement invariance. Given the wide range of infant age in the sample, we controlled for infant age at the assessment of IBQR–VSF in all analyses. Two separate and parallel sets of analyses were conducted to evaluate measurement invariance across racial and income groups. Specifically, a sequence of models was evaluated and compared to test for four increasingly restrictive levels of measurement invariance: configural invariance, metric invariance, scalar invariance, and error variance invariance. Following the procedures outlined by Steenkamp and Baumgartner (1998), we started with evaluating a baseline model for each racial group and then evaluated a series of invariance models step by step. To establish a baseline model, we started with specifying a three-factor confirmatory factor analysis (CFA) model according to the conceptualization of the IBQR–VSF reported by Putnam et al. (2014) and evaluated the model fit. After establishing the baseline model, we evaluated configural invariance by estimating an MGCFA model where all parameters were freely estimated across groups (configural invariance model). Configural invariance would be supported if this model demonstrated adequate fit and all factor loadings were significantly different from zero in both groups.

Next, we tested for metric invariance by evaluating whether factor loadings were equal across racial groups. Specifically, we specified a model with all factor loadings constrained to be equal across groups (metric invariance model) and compared it with the configural invariance model. Full or complete metric invariance would be established if the metric invariance model demonstrates equivalent fit to the configural invariance model. If the metric invariance model demonstrated worse fit than the configural invariance model, it would indicate that at least some of the factor loadings vary across the groups and cannot be constrained to be equal. In the instance that full metric invariance is not established, partial metric invariance can be examined. Partial metric invariance refers to the approach of imposing equality constraints on some but not all of the factor loadings (Byrne, Shavelson, & Muthén, 1989). Researchers have suggested that at least partial metric invariance must be established before evaluating more restrictive invariance (Vandenberg & Lance, 2000).

After establishing metric invariance, we specified a scalar invariance model by adding equality constraints on all of the item intercepts to the previously established metric invariance model.

We tested for scalar invariance by comparing the scalar invariance model and the metric invariance model. Full scalar invariance would be supported if the two models demonstrated equivalent fit. If the scalar invariance model fit significantly worse than the metric invariance model, it would indicate that at least some of the item intercepts differ across groups and thus partial scalar invariance would be evaluated.

Finally, we specified an error variance invariance model by adding additional constraints to the previously established scalar invariance model that error or residual variance for all scale items were equal across groups. We tested for error variance invariance by comparing this model with the previously established scalar invariance model. If the error variance invariance model demonstrated equivalent model fit to the scalar invariance model, it would indicate that error variances were invariant across groups. If the error variance invariance model fit significantly worse than the scalar invariance model, it would suggest that at least some of the error variances differ across groups and thus partial error variance invariance would be evaluated.

Model fit indexes were employed to evaluate the adequacy of fit of each model to the sample data. Several fit indexes, including chi-square (χ^2), the comparative fit index (CFI), and the root mean square error of approximation (RMSEA) were evaluated. A nonsignificant χ^2 indicates good model fit. Greater values of CFI indicate better fit; values greater than .90 indicate adequate fit and values greater than .95 indicate good fit. RMSEA ranges from 0 to 1 with values smaller than .08 indicating adequate fit and values smaller than .05 indicating good fit (Kline, 2011). Because MGCFA models with more parameters constrained to be equal across groups were nested within the less restrictive models where fewer parameters were constrained to be equal, we used chi-square difference ($\Delta\chi^2$) tests between nested models for model comparisons. Given that chi-square difference tests are sensitive to sample size, we also relied on the difference in CFI (ΔCFI) to evaluate measurement invariance as recommended by Cheung and Rensvold (2002). A nonsignificant difference in $\Delta\chi^2$, a ΔCFI no greater than .01 (absolute value), or both would suggest no significant difference in model fit between the nested models and provide support for the aspect of measurement invariance being tested.

Results

To establish a baseline model, we started with examining a basic three-factor CFA model for the IBQR–VSF following Putnam et al. (2014). Specifically, 13 items were specified as loading on a factor representing positive affectivity or surgency, 12 items were specified as loading on a factor representing negative emotionality, and 12 items were specified as loading on a factor representing orienting or regulatory capacity. No cross-loadings or correlated errors were allowed in this model; correlation among the latent factors was allowed. This basic CFA model demonstrated poor fit to the data (Table 2). Consistent with the approach taken by Putnam and Rothbart (2006) in relation to the evaluation of the Very Short Form of the Childhood Behavior Questionnaire (a temperament measure for children 3–8 years of age), a modified CFA model was then conducted with a priori correlated errors allowed between items from the same subscales of the original IBQ. For example, the error terms for all items that originated from the distress to limits subscale were correlated. This modified CFA model (with 38 a priori correlated errors) demonstrated much better model fit than the basic CFA model (Table 2). We considered this model as demonstrating mediocre to adequate fit with CFI above .85 and RMSEA less than

.05 (Kenny, 2014). Some researchers have argued that the conventional goodness-of-fit criteria for CFA models (e.g., CFI > .90) are too restrictive when applied to multifactor instruments with many items (Marsh, Hau, & Wen, 2004), and that personality measures often demonstrate poor fit when evaluated with CFA in part due to the inherent complexity of personality (Hopwood & Donnellan, 2010), as is the case for IBQR–VSF. Although modification indexes pointed to additional correlated errors that could significantly help improve model fit, we decided to keep this modified CFA model with only a priori correlated errors specified to be consistent with original conceptualizations of the IBQR–VSF measure (Putnam et al., 2014). This modified CFA model was used as the baseline model for subsequent measurement invariance tests.

Table 2. Model fit statistics for confirmatory factor analysis models.

	χ^2	*df*	*p*	CFI	RMSEA [90% CI]
Basic CFA model					
Whole sample (*N* = 470)	1951.82	626	<.01	.679	.067 [.064, .071]
White (*n* = 185)	1528.19	626	<.01	.561	.088 [.083, .094]
African American (*n* = 285)	1343.47	626	<.01	.702	.063 [.059, .068]
CFA model with a priori correlated errors					
Whole sample (*N* = 470)	1187.88	588	<.01	.855	.047 [.043, .050]
White (*n* = 185)	1022.30	588	<.01	.789	.063 [.057, .070]
African American (*n* = 285)	944.13	588	<.01	.852	.046 [.041, .051]

Note. CF = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval; CFA = confirmatory factor analysis.

Factor loadings and correlations between latent factors from the basic CFA model and the modified CFA model for the whole sample are presented in Table 3. In both models, factor loadings for all scale items were significantly greater than zero, except for one item. The item "being fed in lap, eager to get away" did not significantly load on the orienting or regulatory capacity factor. Despite the nonsignificant factor loading, this item was still kept in the CFA model for all analyses to be consistent with the currently recommended scoring of the IBQR–VSF measure. Negative emotionality, one latent factor, was negatively associated with orienting or regulatory capacity ($r = –.15$, $p < .05$), another latent factor. The latent factor positive affectivity or surgency was also positively correlated with orienting or regulatory capacity ($r = .77$, $p < .001$). However, the latent factor positive affectivity or surgency was not significantly associated with negative emotionality ($r = –.09$, $p > .05$).

Table 3. Standardized factor loadings from CFA models.

IBQR–VSF Scale/Item	IBQ scale	Basic CFA	CFA with correlated errors
Positive Affectivity/Surgency			
1. Being (un)dressed, squirm away	Activity Level	.25**	.20**
2. When tossed around playfully, laugh	High Pleasure	.50**	.45**
7. Move quickly toward new objects	Approach	.47**	.46**
8. When put into bath water, laugh	Smiling and Laughter	.38**	.41**
13. When placed on back, squirm/turn body	Activity Level	.31**	.28**
14. During a peekaboo game, baby laugh	High Pleasure	.48**	.44**
15. Look up from playing when telephone rings	Perceptual Sensitivity	.44**	.41**
20. Visiting a new place, get excited exploring	Approach	.39**	.38**
21. Smile or laugh when given a toy	Smiling and Laughter	.63**	.68**
26. When hair was washed, baby vocalize	Vocal Reactivity	.19**	.17**

27. Notice sound of an airplane passing overhead	Perceptual Sensitivity	.24**	.23**
36. Make talking sounds when riding in a car	Vocal Reactivity	.47**	.48**
37. When placed in infant or car seat, squirm	Activity Level	.17*	.14*
Negative Emotionality			
3. Tired, show distress	Sadness	.41**	.36**
4. Introduced to unfamiliar adult, cling to parent	Fear	.22**	.14**
9. Time for bed or nap, whimper or sob	Sadness	.50**	.46**
10. After sleeping, cry if someone does not come	Distress to Limitation	.66**	.72**
16. Seem angry when left in the crib	Distress to Limitation	.63**	.68**
17. Startle at a sudden change in body position	Fear	.36**	.36**
22. End of an exciting day, become tearful	Sadness	.34**	.29**
23. Protest being placed in a confining place	Distress to Limitation	.53**	.53**
28. Introduce to unfamiliar adult, refuse to go	Fear	.28**	.22
29. Not able to get attention, cry	Sadness	.69**	.68**
32. Upset when could not get something wanted	Distress to limitation	.57**	.65**
33. Presence of unfamiliar adults, cling to parent	Fear	.27**	.18**
Orienting/Regulatory Capacity			
5. Enjoy being read to	Low Pleasure	.48**	.49**
6. Play with one toy for 5–10 minutes	Duration of Orienting	.31**	.34**
11. Being fed in lap, eager to get away	Cuddliness	−.03	−.07
12. Singing or talking to, soothe immediately	Soothability	.52**	.43**
18. Enjoy hearing the sound of words	Low Pleasure	.50**	.47**
19. Look at pictures in books for 5 minutes or longer	Duration of Orienting	.47**	.52
24. Being held, seem to enjoy himself/herself	Cuddliness	.38**	.31**
25. Showed something to look at, soothe	Soothability	.54**	.48**
30. Enjoy gentle rhythmic activities	Low Pleasure	.32**	.28**
31. Stare at a mobile, crib bumper, picture > 5 minutes	Duration of Orienting	.31**	.32**
34. Rocked or hugged, seem to enjoy	Cuddliness	.41**	.34**
35. Patting or gently rubbing, soothe	Soothability	.53**	.46**
Correlations between factors			
Negative emotionality with surgency		−.05	−.09
Orienting with surgency		.56**	.77**
Orienting with negative emotionality		−.13*	−.15*

Note. $N = 470$. CFA = confirmatory factor analysis; IBQR–VSF = Infant Behavior Questionnaire Revised–Very Short Form; IBQ = Infant Behavior Questionnaire.
* $p < .05$. ** $p < .01$.

Results indicated that the configural invariance model demonstrated adequate fit according to the RMSEA (Table 4). That the CFI of the configural invariance model was lower than .90 is reasonable given the baseline modified CFA model had a similarly low value for the CFI. Although the chi-square difference between the full metric invariance model and the configural invariance model was statistically significant, $\Delta\chi^2(34) = 53.44$, $p < .05$, ΔCFI between these two models was –.004, smaller than .01, indicating that there was invariance in factor loadings across racial groups indicating full metric invariance. A comparison between the full scalar invariance model (factor loadings and all intercepts equal across groups) and the metric invariance model suggested that full scalar invariance was also supported, $\Delta\chi^2(34) = 57.40$, $p < .01$, ΔCFI = –.006, suggesting all item means are invariant across racial groups. Error variance invariance was then evaluated and results indicated that full error variance invariance was also supported given that the full error variance invariance model demonstrated similar fit compared to the full scalar invariance model, $\Delta\chi^2(37) = 78.79$, $p < .01$, ΔCFI = –.009.

Table 4. Model comparisons evaluating measurement invariance across racial groups.

						Model comparisons	
	χ^2	*df*	CFI	RMSEA	Comparison model	$\Delta\chi^2$ (*df*)	ΔCFI
1. Configural invariance	1966.42	1,176	.823	.053 [.049, .058]	—	—	—
2. Full metric invariance	2019.86	1,210	.819	.053 [.049, .057]	1	53.44 (34)	−.004
3. Full scalar invariance	2077.26	1,244	.813	.053 [.049, .057]	2	57.40 (34)	−.006
4. Full error variance invariance	2156.05	1,281	.804	.054 [.050, .058]	3	78.79 (37)	−.009

Note. $N = 470$; White = 185, African American = 285. CFI = comparative fit index; RMSEA = root mean square error of approximation. Confirmatory factor analysis model with a priori correlated errors was used as the baseline model.

We also conducted MGCFA models to examine measurement invariance of IBQR–VSF across income groups, following the same approach we used to evaluate measurement invariance across racial groups. Results are presented in Table 5. Full metric invariance across income groups was supported, as the full metric invariance model demonstrated equivalent fit to the configural invariance model, $\Delta\chi^2(34) = 56.87$, $p < .05$, ΔCFI = –.005. The full scalar invariance model also demonstrated equivalent fit to the full metric invariance model, $\Delta\chi^2(34) = 58.42$, $p < .01$, ΔCFI = –.006, suggesting that full scalar invariance across income groups was also supported. A comparison between the full error variance invariance model and the full scalar invariance model indicated that full error variance invariance was not supported, $\Delta\chi^2(37) = 109.61$, $p < .01$, ΔCFI = –.016. Following recommendations from other researchers, we explored the adequacy of partial error variance invariance (Byrne et al., 1989; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998). Specifically, we reviewed modification indexes to identify which specific error or residual variance should be freed across groups to help improve model fit and establish partial error variance invariance. M*plus* output provides modification indexes that include information about what parameters should be freed across groups to improve model fit, as well as the magnitude of decrease in chi-square statistics associated with freeing each parameter. Following the approach recommended by Byrne and colleagues (1989), we modified the full error variance invariance model by freeing error variances based on modification indexes one by one, starting with the error variance with the largest modification indexes value (i.e., the largest change in chi-square statistics), until the model demonstrated equivalent fit to the scalar invariance model. The criteria for freeing parameters are minimum modification indexes value of 3.84. Partial error variance invariance was established with 3 (out of 37) error variances being freely estimated across groups, $\Delta\chi^2(34) = 67.94$, $p < .01$, ΔCFI = –.007); see Table 6 for the 3 items that varied.

Table 5. Model comparisons evaluating measurement invariance across socioeconomic status groups.

						Model comparisons	
	χ^2	*df*	CFI	RMSEA	Comparison model	$\Delta\chi^2$(*df*)	ΔCFI
1. Configural invariance	1923.03	1,176	.826	.053 [.049, .057]	—	—	—
2. Full metric invariance	1979.90	1,210	.821	.053 [.049, .057]	1	56.87(34)	−.005
3. Full scalar invariance	2038.32	1,244	.815	.053 [.049, .057]	2	58.42(34)	−.006
4. Full error variance invariance	2147.93	1,281	.799	.055 [.051, .059]	3	109.61(37)	−.016
5. Partial error variance invariance	2106.26	1,278	.808	.053 [.049, .057]	3	67.94 (34)	−.007

Note. $N = 454$; high socioeconomic status = 296, low socioeconomic status = 158. CFI = comparative fit index; RMSEA = root mean square error of approximation. Confirmatory factor analysis model with a priori correlated errors was used as the baseline model. The low socioeconomic status group included individuals with income-to-needs ratio lower than 1. These results indicated that metric invariance, scalar invariance, and partial error variance invariance were established. The partial error variance invariance model allowed 3 (out of 37) residual variances to be freely estimated across socioeconomic status groups based on modification indexes.

Table 6. Items in error variances varied across income groups.

IBQR–VSF scale/item	Poor	Not poor
Negative Emotionality		
23. Protest being placed in a confining place	3.13	1.75
29. Not able to get attention, cry	2.08	1.13
Orienting/Regulatory Capacity		
25. Showed something to look at, soothe	1.78	1.05

Note. IBQR–VSF = Infant Behavior Questionnaire Revised–Very Short Form; Item error variances that differ across income groups are reported.

Discussion

The purpose of this study was to evaluate measurement invariance of the IBQR–VSF in a racially and socioeconomically diverse sample of mothers. Results indicated the IBQR–VSF evidenced configural invariance, full metric invariance, full scalar invariance, and full error variance invariance across racial groups. In addition, configural invariance, full metric invariance, full scalar invariance, and partial error variance invariance were established across income groups. Findings regarding configural and full metric invariance suggest that the factor structure of the IBQR–VSF measure and conceptualizations of IBQR–VSF constructs (i.e., positive affectivity or surgency, negative emotionality, and orienting or regulatory capacity) are similar across racial and income groups. That full scalar invariance was established suggests that latent means of the constructs measured by IBQR–VSF can be meaningfully compared across racial and income groups. Full and partial error variance invariance of the IBQR–VSF across racial and income groups suggested that this measure is similarly reliable across groups. Taken together, it appears that White and African American mothers, and both those below and above the poverty line interpret and respond to the IBQR–VSF in similar ways. Thus, it appears that the IBQR–VSF is appropriate for use in racially and economically diverse samples. Future studies that examine differences in mean levels of the IBQR–VSF broadband scales and how these scales are associated with other variables across racial or income groups can be confident that their findings are not likely to be biased due to difference in measurement.

To our knowledge, this is the first work reporting the fit of the three-factor model applied to the IBQR–VSF. Consistent with prior research with the Child Behavior Questionnaire–Very Short Form, the related measure of temperament designed for children age 3 to 8 years, a three-factor model in which error terms for items from the same original subscale (e.g., all items from the fear subscale) were correlated fit the data better than a model in which these error terms were not correlated (Putnam & Rothbart, 2006). This is not surprising, as each of the three broadband scales is composed of items reflecting several distinct dimensions of temperament. For example,

the negative emotionality factor is constructed from items from three different scales from the IBQ–R: Fear, Distress to Limits, and Sadness. Items from the same original scale are likely more highly correlated with one another than with items from the other scales. Although the model with correlated errors fit better than the model with uncorrelated errors, the fit was marginal (CFI lower than .90 but RMSEA was between .05 and .08), suggesting the three-factor structure does not adequately capture the dimensions that make up a child's temperament (Allan, Lonigan, & Wilson, 2013). In future research, the factor structure of the IBQR–VSF should be examined via CFA in larger samples. Given it is likely that many researchers are using the IBQR–VSF in its current form, we opted to maintain the original factor structure when investigating invariance.

The pattern of intercorrelations among the scale scores is also notable. That negative emotionality and positive affectivity or surgency were negatively but not significantly associated ($r = –.09$) is inconsistent with prior research with the IBQ–R ($r = .16$; Gartstein & Rothbart, 2003), and counter to the argument that children are generally predisposed to reactivity, both positive and negative in nature. That positive affect or surgency and orienting or regulatory capacity were so highly positively correlated ($r = .77$) is inconsistent with prior research ($r = .25$; Gartstein & Rothbart, 2003). The difference could be a function of infant age, such that the two constructs become more distinct as infants mature given the majority of infants in this sample were 6 months or younger, but in the other sample, the majority were 6 months or older. This strong association likely contributed to the marginal model fit. That negative emotionality and orienting or regulatory capacity are negatively associated ($r = –.15$) is consistent with prior research, although the magnitude is smaller ($r = –.30$; Gartstein & Rothbart, 2003), and might indicate that negative reactivity and regulation can be rated by parents independent of one another.

Limitations of this research include our inability to examine invariance between additional racial and ethnic groups, including Hispanic and Asian parents, and to test for invariance among four possible race and income groups due to limited cell size. In addition, our combined sample comes from studies that used different response methods. Future research should address these issues and also examine invariance in predictive validity of the IBQR–VSF in relation to child, parent, and family outcomes. Such research should include fathers also.

In conclusion, this study provides evidence for measurement invariance of the IBQR–VSF across African American and White mothers and poor versus not poor mothers. The results suggest the IBQR–VSF is a parental report measure of infant temperament that is appropriate for use in diverse samples. This provides partial support for use of the IBQR–VSF in large-scale developmental and epidemiological research, as suggested by Putnam et al. (2014), although additional work on invariance in other racial groups is warranted.

Acknowledgments

We are grateful to the families who participated and to our dedicated project staff who oversaw and completed data collection.

Funding

This research was supported by R01HD061010 and R01HD058578. The content is solely the responsibility of the authors and does not represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

References

1. Allan, N. P., Lonigan, C. J., & Wilson, S. B. (2013). Psychometric evaluation of the Children's Behavior Questionnaire–Very Short Form in preschool children using parent and teacher report. *Early Childhood Research Quarterly*, *28*, 302–313. doi:10.1016/j.ecresq.2012.07.009 [Crossref], [Web of Science ®], [Google Scholar]

2. Bernstein, R. E., Tenedios, C. M., Laurent, H. K., Measelle, J. R., & Ablow, J. C. (2014). The eye of the begetter: Predicting infant attachment disorganization from women's prenatal interpretations of infant facial expressions. *Infant Mental Health Journal*, *35*, 233–244. doi:10.1002/imhj.21438 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

3. Bridgett, D. J., Gartstein, M. A., Putnam, S. P., Lance, K. O., Iddins, E., Waits, R., & … Lee, L. (2011). Emerging effortful control in toddlerhood: The role of infant orienting/regulation, maternal effortful control, and maternal time spent in caregiving activities. *Infant Behavior & Development*, *34*, 189–199. doi:10.1016/j.infbeh.2010.12.008 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

4. Bridgett, D. J., Gartstein, M. A., Putnam, S. P., McKay, T., Iddins, E., Robertson, C., & … Rittmueller, A.(2009). Maternal and contextual influences and the effect of temperament development during infancy on parenting in toddlerhood. *Infant Behavior & Development*, *32*, 103–116. doi:10.1016/j.infbeh.2008.10.007 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

5. Burney, R. V., & Leerkes, E. M. (2010). Links between mothers' and fathers' perceptions of infant temperament and coparenting. *Infant Behavior & Development*, *33*, 125–135. doi:10.1016/j.infbeh.2009.12.002 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

6. 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. doi:10.1037/0033-2909.105.3.456 [Crossref], [Web of Science ®], [Google Scholar]

7. Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*, 155–175. doi:10.1177/0022022102250225 [Crossref], [Web of Science ®], [Google Scholar]

8. Calkins, S. D., Dedmon, S. E., Gill, K. L., Lomax, L. E., & Johnson, L. M. (2002). Frustration in infancy: Implications for emotion regulation, physiological processes, and

temperament. *Infancy*, *3*, 175–197. doi:10.1207/S15327078IN0302_4 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

9. Calkins, S. D., Hungerford, A., & Dedmon, S. E. (2004). Mothers' interactions with temperamentally frustrated infants. *Infant Mental Health Journal*, *25*, 219–239. doi:10.1002/imhj.20002 [Crossref], [Web of Science ®], [Google Scholar]

10. Carey, W. B., & McDevitt, S. C. (1978). Revision of the Infant Temperament Questionnaire. *Pediatrics*, *61*, 735–739. [PubMed], [Web of Science ®], [Google Scholar]

11. Chen, J. (2012). Maternal alcohol use during pregnancy, birth weight and early behavioral outcomes. *Alcohol and Alcoholism*, *47*, 649–656. doi:10.1093/alcalc/ags089 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

12. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

13. Cole, P. M., & Tan, P. Z. (2007). Emotion socialization from a cultural perspective. In J. E. Grusec & P. D.Hastings (Eds.), *Handbook of socialization: Theory and research* (pp. 516–542). New York, NY: Guilford Press. [Google Scholar]

14. Crockenberg, S., & Leerkes, E. (2003). Infant negative emotionality, caregiving, and family relationships. In A. C. Crouter & A. Booth (Eds.), *Children's influence on family dynamics: The neglected side of family relationships* (pp. 57–78). Mahwah, NJ: Erlbaum. [Google Scholar]

15. Dennis, P. A., & Halberstadt, A. G. (2013). Is believing seeing? The role of emotion-related beliefs in selective attention to affective cues. *Cognition and Emotion*, *27*, 3–20. doi:10.1080/02699931.2012.680578 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

16. Garrett-Peters, P., Mills-Koonce, R., Adkins, D., Vernon-Feagans, L., & Cox, M. (2008). Early environmental correlates of maternal emotion talk. *Parenting: Science and Practice*, *8*, 117–152. doi:10.1080/15295190802058900 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

17. Gartstein, M. A., Putnam, S. P., & Rothbart, M. K. (2012). Etiology of preschool behavior problems: Contributions of temperament attributes in early childhood. *Infant Mental Health Journal*, *33*, 197–211. doi:10.1002/imhj.21312 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

18. Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the revised infant behavior questionnaire. *Infant Behavior & Development*, *26*(1), 64–86. doi:10.1016/S0163-6383(02)00169-8 [Crossref], [Web of Science ®], [Google Scholar]

19. Gartstein, M. A., Slobodskaya, H. R., Putnam, S. P., & Kinsht, I. A. (2009). A cross-cultural study of infant temperament: Predicting preschool effortful control in the United States of America and Russia. *European Journal of Developmental Psychology*, *6*, 337–364. doi:10.1080/17405620701203846 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

20. Ghera, M. M., Hane, A. A., Malesa, E. E., & Fox, N. A. (2006). The role of infant soothability in the relation between infant negativity and maternal sensitivity. *Infant Behavior & Development*, *29*, 289–293.doi:10.1016/j.infbeh.2005.09.003 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

21. Harrell-Williams, L. M., Raines, T. C., Kamphaus, R. W., & Dever, B. V. (2015). Psychometric analysis of the BASC–2 Behavioral and Emotional Screening System (BESS) student form: Results from high school student samples. *Psychological Assessment*, *27*, 738–743. doi:10.1037/pas0000079 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

22. Holly, L. E., Little, M., Pina, A. A., & Caterino, L. C. (2015). Assessment of anxiety symptoms in school children: A cross-sex and ethnic examination. *Journal of Abnormal Child Psychology*, *43*, 297–309. doi:10.1007/s10802-014-9907-4 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

23. Hopwood, C., & Donnellan, M. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

24. Kenny, D. A. (2014). Measuring model fit. Retrieved from http://davidakenny.net/cm/fit.htm [Google Scholar]

25. Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York NY: Guilford. [Google Scholar]

26. Lahey, B. B., Van Hulle, C. A., Keenan, K., Rathouz, P. J., D'Onofrio, B. M., Rodgers, J. L., & Waldman, I. D. (2008). Temperament and parenting during the first year of life predict future child conduct problems. *Journal of Abnormal Child Psychology*, *36*, 1139–1158. doi:10.1007/s10802-008-9247-3 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

27. Leerkes, E. M., & Burney, R. V. (2007). The development of parenting efficacy among new mothers and fathers. *Infancy*, *12*(1), 45–67.doi:10.1111/j.1532-7078.2007.tb00233.x [Crossref], [Web of Science ®], [Google Scholar]

28. Leerkes, E. M., Supple, A. J., O'Brien, M., Calkins, S. D., Haltigan, J. D., Wong, M. S., & Fortuna, K. (2015). Antecedents of maternal sensitivity during distressing tasks: Integrating attachment, social information processing, and psychobiological perspectives. *Child Development*, *86*, 94–111. doi:10.1111/cdev.12288 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

29. Leve, L. D., Scaramella, L. V., & Fagot, B. I. (2001). Infant temperament, pleasure in parenting, and marital happiness in adoptive families. *Infant Mental Health Journal*, *22*, 545–558. doi:10.1002/imhj.1017 [Crossref], [Web of Science ®], [Google Scholar]

30. Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341. [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

31. Matsumoto, D. (1989). Face, culture, and judgments of anger and fear: Do the eyes have it? *Journal of Nonverbal Behavior*, *13*, 171–188. doi:10.1007/BF00987048 [Crossref], [Web of Science ®], [Google Scholar]

32. Matsumoto, D. (1993). Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. *Motivation and Emotion*, *17*, 107–123. doi:10.1007/BF00995188 [Crossref], [Web of Science ®], [Google Scholar]

33. McCoy, D. C., Raver, C. C., Lowenstein, A. E., & Tirado-Strayer, N. (2011). Assessing self-regulation in the classroom: Validation of the BIS-11 and the BRIEF in low-income, ethnic minority school-age children. *Early Education and Development*, *22*, 883–906. doi:10.1080/10409289.2010.508371 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

34. Mehall, K. G., Spinrad, T. L., Eisenberg, N., & Gaertner, B. M. (2009). Examining the relations of infant temperament and couples' marital satisfaction to mother and father involvement: A longitudinal study. *Fathering*, *7*(1), 23–48. doi:10.3149/fth.0701.23 [Crossref], [PubMed], [Google Scholar]

35. Nelson, J. A., Leerkes, E. M., O'Brien, M., Calkins, S. D., & Marcovitch, S. (2012). African American and European American mothers' beliefs about negative emotions and emotion socialization practices. *Parenting: Science and Practice*, *12*, 22–41. doi:10.1080/15295192.2012.638871 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

36. Oddi, K. B., Murdock, K. W., Vadnais, S., Bridgett, D. J., & Gartstein, M. A. (2013). Maternal and infant temperament characteristics as contributors to parenting stress in the first year postpartum. *Infant and Child Development*, *22*, 553–579. doi:10.1002/icd.1813 [Crossref], [Web of Science ®], [Google Scholar]

37. Parade, S. H., & Leerkes, E. M. (2008). The reliability and validity of the Infant Behavior Questionnaire-Revised. *Infant Behavior & Development*, *31*, 637–646. doi:10.1016/j.infbeh.2008.07.009 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

38. Parker, A. E., Halberstadt, A. G., Dunsmore, J. C., Townley, G., Bryant, A., Thompson, J. A., & Beale, K. S.(2012). “Emotions are a window into one's heart”: A qualitative analysis of parental beliefs about children's emotions across three ethnic groups: I. Introduction. *Monographs of the Society for Research in Child Development*, *77*(3), 1–109. doi:10.1111/j.1540-5834.2012.00677.x [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

39. Planalp, E. M., Braungart-Rieker, J. M., Lickenbrock, D. M., & Zentall, S. R. (2013). Trajectories of parenting during infancy: The role of infant temperament and marital adjustment for mothers and fathers. *Infancy*, *18*(Suppl. 1), E16–E45. doi:10.1111/infa.12021 [Crossref], [Web of Science ®], [Google Scholar]

40. Putnam, S. P., Helbig, A. L., Gartstein, M. A., Rothbart, M. K., & Leerkes, E. (2014). Development and assessment of Short and Very Short Forms of the Infant Behavior Questionnaire–Revised. *Journal of Personality Assessment*, *96*, 445–458. doi:10.1080/00223891.2013.841171 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

41. Putnam, S. P., & Rothbart, M. K. (2006). Development of Short and Very Short Forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*, 102–112. doi:10.1207/s15327752jpa8701_09 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

42. Rothbart, M. K. (1981). Measurement of temperament in infancy. *Child Development*, *52*, 569–578. doi:10.2307/1129176 [Crossref], [Web of Science ®], [Google Scholar]

43. Rothbart, M. K. (1986). Longitudinal observation of infant temperament. *Developmental Psychology*, *22*, 356–365. doi:10.1037/0012-1649.22.3.356 [Crossref], [Web of Science ®], [Google Scholar]

44. Rothbart, M. K., & Bates, J. E. (2006). Temperament. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 99–166). Hoboken, NJ: Wiley. [Google Scholar]

45. Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210–222. doi:10.1016/j.hrmr.2008.03.003 [Crossref], [Web of Science ®], [Google Scholar]

46. Schoppe-Sullivan, S. J., Mangelsdorf, S. C., Brown, G. L., & Sokolowski, M. S. (2007). Goodness-of-fit in family context: Infant temperament, marital quality, and early coparenting behavior. *Infant Behavior & Development*, *30*, 82–96. doi:10.1016/j.infbeh.2006.11.008 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

47. Solmeyer, A. R., & Feinberg, M. E. (2011). Mother and father adjustment during early parenthood: The roles of infant temperament and coparenting relationship quality. *Infant*

Behavior & Development, *34*, 504–514. doi:10.1016/j.infbeh.2011.07.006 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

48. Steenkamp, J. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90. doi:10.1086/209528 [Crossref], [Web of Science ®], [Google Scholar]

49. Thomas, A., & Chess, S. (1977). *Temperament and development*. Oxford, UK: Brunner/Mazel. [Google Scholar]

50. Trent, L. R., Buchanan, E., Ebesutani, C., Ale, C. M., Heiden, L., Hight, T. L., & … Young, J. (2013). A measurement invariance examination of the Revised Child Anxiety and Depression Scale in a Southern sample: Differential item functioning between African American and Caucasian youth. *Assessment*, *20*, 175–187. [Crossref], [PubMed], [Web of Science ®], [Google Scholar]

51. van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. doi:10.1080/17405629.2012.686740 [Taylor & Francis Online], [Web of Science ®], [Google Scholar]

52. 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*(1), 4–69. doi:10.1177/109442810031002 [Crossref], [Web of Science ®], [Google Scholar]

53. Willoughby, M. T., Stifter, C. A., & Gottfredson, N. C. (2015). The epidemiology of observed temperament: Factor structure and demographic group differences. *Infant Behavior & Development*, *39*, 21–34. doi:10.1016/j.infbeh.2015.02.001 [Crossref], [PubMed], [Web of Science ®], [Google Scholar]