Published in final edited form as:
Infant Behav Dev. 2015 May ; 39: 21–34. doi:10.1016/j.infbeh.2015.02.001.

The Epidemiology of Observed Temperament: Factor Structure and Demographic Group Differences

Michael T. Willoughby[1], **Cynthia A. Stifter**[2], **Nisha C. Gottfredson**[3], and **The Family Life Project Investigators**

[1]Education & Workforce Development, RTI International

[2]Human Development & Family Studies, Pennsylvania State University

[3]Center for Developmental Science, University of North Carolina at Chapel Hill

Abstract

This study investigated the factor structure of observational indicators of children's temperament that were collected across the first three years of life in the Family Life Project (N = 1205) sample. A four-factor model (activity level, fear, anger, regulation), which corresponded broadly to Rothbart's distinction between reactivity and regulation, provided an acceptable fit the observed data. Tests of measurement invariance demonstrated that a majority of the observational indicators exhibited comparable measurement properties for male vs. female, black vs. white, and poor vs. not-poor children, which improved the generalizability of these results. Unadjusted demographic group comparisons revealed small to moderate sized differences (Cohen ds = |.23 – .42|) in temperamental reactivity and moderate to large sized differences (Cohen ds = −.64 – −.97) in regulation. Collectively, demographic variables explained more of the variation in regulation (R^2 = .25) than in reactivity (R^2 = .02 – .06). Follow-up analyses demonstrated that race differences were substantially diminished in magnitude and better accounted for by poverty. These results help to validate the distinction between temperamental reactivity and regulation using observational indicators.

Keywords

temperament; reactivity; regulation; latent variable; confirmatory factor analysis

Contact Information: Correspondence should be directed to Michael Willoughby, RTI International, Hobbs #349, Research Triangle Park, NC 27709-2194 or preferably mwilloughby@rti.org.

1.0 The Epidemiology of Observed Temperament: Factor Structure and Demographic Group Differences

Numerous models and definitions of temperament exist (Goldsmith et al., 1987; Rothbart, Derryberry, & Posner, 1994; Strelau, 1994), a consensus definition characterizes temperament as individual differences in behavioral tendencies that are evident early in life and reflect early biological predispositions that are shaped by contextual experience (Rothbart & Bates, 1998). Scholars from a wide range of disciplines are empirically interested in the construct of temperament. For example, scientists who study prenatal development routinely consider temperament as an early outcome that is associated with early exposure to toxicants, (non)prescription drugs, and general stressors (Blair, Glynn, Sandman, & Davis, 2011; Grey, Davis, Sandman, & Glynn, 2013; Mayes, 2002; Richardson, Goldschrmidt, & Willford, 2008; Schuetze, Molnar, & Eiden, 2012; Weiss, Jonn-Seed, & Harris-Muchell, 2007). Cognitive neuroscientists have characterized temperament as “model area of study” for questions focused on the inter-relations of cognitive and emotional functioning (Bell & Wolfe, 2004; Henderson & Wachs, 2007; Wolfe & Bell, 2004). Clinical psychologists and psychiatrists are concerned with the prognostic value of temperament to forecast emergent psychopathology in early and middle childhood (Bijttebier & Roeyers, 2009; Martel, 2009; Muris & Ollendick, 2005; Nigg, 2006; Rapee & Coplan, 2010).

Parent-report inventories remain the most widely utilized method for measuring temperament (Garstein, Bridgett, & Low, 2012). The popularity of parent-reports of temperament is due both to their ease of administration and acknowledgment of parents’ unique vantage of their children’s behavioral tendencies across time and settings. Due to the ease of administration and scoring, large-scale studies have typically relied exclusively on parent report questionnaires to measure temperament (Henrichs et al., 2009; Kaplan et al., 2014; Sanson, Prior, Garino, Oberklaid, & Sewell, 1987; Wessman et al., 2012). Studies involving parent reported inventories of temperament have provided empirical support for sub-dividing the construct of temperament into two broad domains—reactivity and regulation (Putnam & Stifter, 2008). Whereas reactivity has been defined as “the speed, strength, and valence [positive or negative] of an individual’s characteristic response to stimulation”, regulation has been defined as “behaviors the individual uses to control their behavioral and emotional reactions to sources of both positive and negative stimulation” (Henderson & Wachs, 2007; p. 400).

Despite their ease of use, parent reports of temperament have been criticized because they tend to be weakly associated with observed temperamental behaviors (Forman et al., 2003; Seifer, Sameroff, Dickstein, Schiller, & Hayden, 2004; Vaughn, Bradley, Joffe, Seifer, & Barglow, 1987; Vaughn, Taraldson, Crichton, & Egeland, 1981). Characteristics of parents (personality, psychopathology, stress) and parent-child interaction quality have been identified as potential threats to the validity of parent reports of temperament (Forman et al., 2003; Leerkes & Crockenberg, 2003; Mebert, 1991; Parade & Leerkes, 2008; Sameroff, Seifer, & Elias, 1982). Conversely, the novelty and artificiality of observational tasks that are used to elicit temperamental behaviors, the variations in time scale in which behaviors are observed (seconds-minutes for observational measures vs. days/weeks for ratings), and contextual differences all represent alternative explanations for the weak associations between parent reported and observed temperament. Regardless of the specific reasons, the lack of a strong relation between parent report and observed temperament undermines conventional measurement wisdom, which implies that there is scientific value in aggregating information across informant and methods (Campbell & Fiske, 1959; Podsakoff, MacKenzie, & Podsakoff, 2012). Specifically, to the extent that parent reports and observed temperament are weakly associated, there exists little shared variation to define the latent constructs of interest.

Data collectors who conduct laboratory or in-home visits represent another potential source of information on children's temperament that are not subject to the same concerns related to parent rated temperament (e.g., Matheny, 1983). Data collector's global impressions of children's temperament complements information that is obtained from microsocial (e.g., second by second) coding of children's responses to challenge tasks (Carnicero, Perez-Lopez, Del Carmen, Salinas, & Martinez-Fuentes, 2000; Gagne, Van Hulle, Aksan, Essex, & Goldsmith, 2011; Stifter & Corey, 2001). Indeed, we previously demonstrated the merits of using home visitors reports of temperament in the same sample as will be used in the current study (Stifter, Willoughby, Towe-Goodman, & Investigators, 2008). Those results demonstrated that while there was convergence across parents, home visitors, and observation measures of infant positivity, only home visitor and observational measures converged with respect to the measurement of infant negativity. Presumably, home visitor impressions were influenced by their administration of challenge tasks, while parent reports took into account other sources of information.

The first objective of this study was to test whether home visitors ratings and multiple observational (including performance-based) tasks might be used to build latent constructs of three subdomains of temperamental reactivity (fear, anger, activity level), as well as a single domain of regulation. This builds on a small number of recent studies that have begun to systematically investigate the structure and stability of temperament in early and middle childhood using exclusively observational measures (Durbin, Hayden, Klein, & Olino, 2007; Dyson, Olino, Durbin, Goldsmith, & Klein, 2012; Gagne et al., 2011; Kotelnikova, Olino, Mackrell, Jordan, & Hayden, 2013). Our proposed work most closely resembles that of Dyson et al. (2012) and Kotelnikova et al. (2013), both of whom utilized factor analytic methods to test the structure of observed temperament in early and middle childhood, respectively. While the current study differed in important ways from those studies—e.g., sample acquisition (representative vs. convenience), study design (longitudinal vs. cross-section), child age (birth – 3 years vs. early or middle childhood), nature and scope of coding systems, and the nature of constructs considered—the guiding premise of the current study was the same as those previous studies, namely to provide a vantage of the structure of temperament that was independent of parent reports.

A prevailing assumption in the literature is that observational tasks designed to elicit temperamental reactivity and regulation work equally well for all populations. Tests of measurement invariance provide one means of evaluating this assumption. Measurement invariance involves testing a sequence of models that impose increasingly stringent requirements regarding the measurement equivalence of a set of indicators of a latent construct across groups (Meredith, 1993). Here, we exploit the characteristics of the current sample (large N, over-sampling of low income and African American families) to test the measurement invariance of observational measures of temperamental reactivity and regulation as a function of child sex, race, and household poverty status. To be clear, testing whether observational indicators of temperament cohere in comparable ways across children of demographic groups provides one means of evaluating the generalizability of the organizing distinction between reactivity and regulation. We are unaware of any previous studies that have tested these questions using observational data, though this approach has been utilized with parent- and self-report measures (Carter, Briggs-Gowan, Jones, & Little, 2003; Kim, Brody, & Murry, 2003; Zimprich & Mascherek, 2012).

To the extent that temperamental constructs of reactivity and regulation exhibit at least partially strong invariance across groups, this ensures that any observed group differences are meaningful. A tertiary objective of this study was to evaluate group differences as a function of child sex, race, and household poverty status. This objective should not be misconstrued as either an implicit or explicit assumption that demographic factors are causally related to emerging individual differences in temperamental reactivity or regulation. However, given that there already exists a literature that has focused on mean differences in temperament, primarily using parent-report measures, we sought to provide complementary information using observational measures. Evaluating latent mean level differences was a natural byproduct of our primary focus on testing for measurement invariance of temperamental constructs for demographically-defined subgroups of children.

With respect to gender differences, Else-Quest and colleagues (2006) conducted a meta-analysis of three broad dimensions of temperament—negative affect, surgency, and effortful control—which correspond to dimensions of reactivity and regulation considered here. Moderate gender differences were evident for effortful control (girls > boys) and surgency (boys > girls in activity level and impulsivity; girls > boys for approach, positive mood, shyness), while negligible differences were evident for negative affect. Importantly, virtually all of the studies that were included in the Else-Quest et al (2006) meta-analysis utilized parent report data, which were susceptible to potential stereotypes or biases about typical gendered behavior. In addition, an earlier meta-analysis reported small gender differences in activity level (boys > girls) that were observable infancy using parent report and objective measures (Campbell & Eaton, 1999).

Far fewer studies have evaluated group differences in temperament as a function of socioeconomic status (SES), and the available evidence is mixed. While a few early studies reported no associations (Maziade, Boudreault, Thivierge, Caperaa, & Cote, 1984; Persson-Blennow & Mcneil, 1981), others indicated that children from lower income backgrounds exhibited more “difficult” (less desirable) temperaments (Jansen et al., 2009; Oberklaid, Prior, Sanson, Sewell, & Kyrios, 1990; Sameroff et al., 1982). Many of these studies relied on historically dated conceptualizations of temperament. Although we are not aware of studies that have investigated SES differences in aspects of temperamental reactivity, there is evidence that children from low income households exhibit poorer regulation, as indicated by measures of effortful control, executive control, and executive function (Hackman & Farah, 2009; Li-Grining, 2007; Raver, 2012; Zalewski et al., 2012).

Although numerous studies have considered cross-cultural differences in temperament (Oakland & Lu, 2006; Rubin et al., 2006; Slobodskaya, Gartstein, Nakagawa, & Putnam, 2013; Zhou, Lengua, & Wang, 2009), relatively few studies have tested for race differences in temperamental reactivity and regulation in US samples. We are only aware of one study, which reported that African American toddlers were more positively and negatively reactive than European American toddlers (Calkins, Dedmon, Gill, Lomax & Johnson, 2002). Given the confounding of poverty and race in the US, any evidence of group differences in temperament would be difficult to resolve. Nonetheless, given that African American children were over-sampled in our study, we considered measurement invariance and group differences as a function of race and further qualified whether any apparent group differences were better accounted for by poverty.

In sum, the overarching objective of this study was to test whether observational measures, obtained across the first three years of life, could be used to build latent constructs of temperamental reactivity (activity, fear, anger) and regulation, with an interest in the relative contributions of specific indicators, as well as the correlations between latent constructs. In addition, we sought to test whether the fit of our hypothesized measurement model was comparable for children of differing sex, race, and household poverty status.

2.0 Methods

2.1 Participants

The Family Life Project (FLP) is a prospective longitudinal study of families residing in six low-wealth counties in Eastern North Carolina and Central Pennsylvania (3 counties per state) that were selected to be indicative of the Black South and Appalachia, respectively. Complex sampling procedures were employed to recruit a representative sample of 1292 children whose families resided in one of the six counties at the time of the child’s birth. Low-income families in both states and, in North Carolina, African American families were over-sampled; however, through the use of weighted analyses, all of our inferences generalize back to the 6-county study area as if participants were selected using simple random sampling. Full details of the sampling plan and study design appear elsewhere (Vernon-Feagans, Cox, & Investigators, 2013).

The current study utilized data from the 6, 15, 24, and/or 36 month home visits. Two home visits, typically completed within 2-weeks of each other, were conducted at the 6, 24, and 36 month assessments, while only one home visit was conducted at the 15 month assessment. Visits consisted of activities that included parents and children interacting together (e.g., free play interaction, book reading) and separately (e.g., child completion of LAB TAB tasks, Bayley exam, health screening; parent: interviews, questionnaire completion). Two home visitors were involved in every home visit; each home visitor provided independent ratings of children’s behavior following each visit.

Although the majority of families and children participated in their home visit within 2-weeks of the target age (the median child ages for all available children who participated in the 6, 15, 24, and 36 month assessments were 7.6, 15.4, 24.3, 36.6 months, respectively), some families completed their visits much later than the target age. Because the temperament tasks were selected to elicit behaviors at a particular age, we imposed age cutoffs. Specifically, children had to be less than 9 months old at their target 6 month visit (N=935/1202), less than 18 months at their target 15 month visit (N=1092/1169), less than 30 months at their target 24 month visit (N=1095/1142), and less than 39 months at their target 36 month visit (N=935/1118) for their temperament data to be utilized from a given assessment. Of the total sample of N=1292 children, N=1205 (93%) contributed temperament data at one of the four (i.e., age 6, 15, 24, 36 month) assessments (58% contributed data from all 4 assessments, 25% from 3 assessments, 12% from 2 assessments, and 6% from one assessment; percentages do not sum to 100 due to rounding). Participating families and children (N = 1205) did not differ from non-participating families and children (N = 87) with respect to being recruited into the low income stratum (78% vs. 70% poor, *p* = .09), primary caregiver educational status at study enrollment (80% vs. 83% with a high school degree/GED or beyond, *p* = .53), child gender (51% vs. 47% male, *p* = .47), or child race (43% vs. 39% African American, *p* =.51). However, participating families were disproportionally from Pennsylvania (41% vs. 29%, *p* = .02).

2.2 Procedures

We utilized both observer ratings and direct assessments in order to derive latent variable representations of temperamental reactivity and regulation. As reactivity changes with development of regulatory functions, indicators of reactivity were drawn from the age 6, 15, and 24 month assessments. In contrast, indicators of regulation were drawn from the age 24 and 36 month assessments.

2.3 Measures

2.31 Reactivity: Home Visitor Ratings (6–24 Months)—After each home visit, both home visitors independently made global ratings of children's behavior (see Stifter & Corey, 2001 for precedent) using items that were adapted from the Infant Behavior Record (IBR; Bayley, 1969). Hence, at the 6, and 24 month assessments, four independent ratings (two ratings per visit for two visits) were available. At the 15 month assessment, two independent ratings (two rating for the single home visit) were available. Each item was rated on a 9 point Likert scale. Home visitor ratings of the *Amount of gross bodily movement* item (Likert anchors: 1= "Stays quietly in one place, with practically no self-initiated movement" to 9 = "Hyperactive, cannot be quieted for sedentary tasks") were used as an indicator of activity level at the 6, 15, and 24 month assessments (αs = .79, .74, and .80, respectively). Home visitor ratings of the *Reaction to the new or strange* item (Likert anchors: 1 = "Accepts the entire situation with no evidence of fear, caution or inhibition of actions" to 9 = "Strong indication of fear of the strange, to the extent that he cannot be brought to play or respond to the examiner or tasks") were used as an indicator of fear reactivity at the 15 and 24 month assessments (αs = .74 and .75, respectively). Home visitor ratings of the *Irritability* item (Likert anchors: 1 = "No irritability, infant passively responses to all stimulation" to 9 = "Irritable to all degrees of stimulation encountered throughout the home visit") were used as an indicator of anger reactivity at the 15 and 24 month assessments (αs = .79 and .78, respectively).

2.32 Reactivity: Observed Activity Level (6 Months)—Parents and children participated in a free play interaction at the 6 month assessment and videotaped interactions were coded to assess multiple dimensions of global parenting and child behaviors across each interaction. The current study made use of the child activity level code from the 6 month visits. Ratings for activity level (extent to which the child is motorically active during the observation) were made on a 5-point scale ranging from 1= "Not at all characteristic" to 5 = "Highly characteristic". Coders were trained and certified as reliable prior to coding. A minimum of 30% of all observations were double coded throughout the coding period and discrepancies in coding were resolved by conferencing. Coding pairs exhibited acceptable inter-rater reliability for the activity code at 6 months (K = .63).

2.33 Reactivity: Challenge Tasks (15 & 24 Months)—Infants participated in two tasks that were drawn from the LAB-TAB (Goldsmith & Rothbart, 1996). Each task was videotaped and coded off-line by trained research assistants. The *mask task*, which was designed to elicit individual differences in fear reactivity, was administered at the 15 and 24 month home visits. During the task, the home visitor would put on an unusual mask and move her head slowly from side to side while calling the child's name. A total of four masks presented, for approximately 10 seconds each. The *toy removal task*, which was designed to elicit anger reactivity, was administered at the 15 and 24 month assessments (Stifter & Braungart, 1995). During the task, infant and mother play with an interesting toy for 90 seconds after which the mother removes the toy and places it on her seat. The mother then moves away to converse with the home visitor for 2 minutes. The mother than returns the toy and resumes speaking with the home visitor. After one minute the mother returns and resumes interacting with the child or soothes him/her if needed.

Both tasks were subjected to second-by-second coding using the Better Coding Approach software (Danville, Pennsylvania). Three levels of task-related negative reactivity were coded: low negative reactivity, moderate negative reactivity, and high negative reactivity. A weighted negative reactivity composite was calculated for each task (i.e., the average proportion of task time spent exhibiting low = 1, moderate = 2, and high = 3 levels of reactivity). All coders were trained to achieve at least .75 (Cohen's kappa) reliability on the reactivity coding. Subsequent inter-rater reliability was calculated on 15% of cases using kappa coefficients (Ks = .96 and .93 for the mask task at 15 and 24 months; Ks = .88 and .90 for the toy removal task at 15 and 24 months).

2.34 Regulation: Home Visitor Ratings (24 & 36 months)—Home visitor ratings on the *Tendency to persist in attention to any one objective, person, or activity, aside from attaining a goal* (Likert anchors: 1 = "Fleeting attention span" to 9 = "Long-continued absorption in a toy, activity or person") were used as an indicator of regulation at the 24 month assessment (α = .75) and on the *Task Persistence - Degree of on-task behavior, persistence in the face of frustration* item (Likert anchors: 1 = "Does not stay on task, short attention span" to 9 = "Shows high level of task persistence) was used as an indicator at the 36 month assessment (α = .80).

2.35 Regulation: Observed Persistence (24 & 36 Months)—Parents and children participated in a joint puzzle-solving interaction at the 24 and 36 month assessments. Videotaped interactions were coded to assess multiple dimensions of global parenting and child behaviors across each interaction. The child persistence codes from the 24 and 36 month visits were examined in the present study. Ratings for child persistence were made on a 7-point scale ranging from 1 = "Very low" (e.g., Child actively tries to avoid the task and spends as little time as s/he can get away with doing the tasks at all) to 7 = "Very high" (e.g., Child is persistent and works at each task with an apparent goal of getting correct solutions until the problem is solved or exhaustively approximated). The persistence code was intended to reflect the child problem-solving efforts regardless and independent of the degree to which the parent was instrumental in facilitating the child's persistence. Coders double coded 30% of all observations and exhibited acceptable inter-rater reliability for the persistence code at the 24 and 36-month assessments (Ks = .75 and .76, respectively).

2.36 Regulation: Direct Assessments of Inhibitory Control and Attention Shifting (24 & 36 Months)

Children completed one inhibitory control and one attention shifting task at the 24 month visit. The *snack delay task,* adopted from Kochanska and colleagues (2000), was used an indicator of inhibitory control (delay of gratification). In this task, the experimenter placed a desirable snack (small candy or cracker) underneath a transparent container. The child was told to wait until the experimenter rang a bell before retrieving the snack. Each child completed four trials: a 10-second delay followed by 20-, 30-, and 15-second delay trials. Each response was coded as no wait (0), partial wait (1), or full wait (2), and the mean score across trials indicated delay of gratification. The *reverse categorization task*, adopted from Carlson and colleagues (2004), was used as an indicator of attention shifting. In this task, children initially sort small/big sized blocks into corresponding sized buckets but are subsequently asked to sort small blocks into big buckets and vice versa. Three trials (practice, preswitch, switch) of six blocks were administered. Children were designated as passing a trial if they correctly sorted four of six blocks. The sum of passed trials (0–3) was the dependent variable.

Children completed three inhibitory control and one attention shifting task at the 36 month assessment. Full task descriptions, administration and scoring procedures, and psychometric properties of individual tasks were elaborated elsewhere (Willoughby, Blair, Wirth, Greenberg, & Investigators, 2010). Briefly, the *Silly Sounds Stroop* task (inhibitory control) asks children to make a barking sound when show pictures of a cat and a meow sound when shown pictures of a dog. The *Spatial Conflict* task (inhibitory control) is a Simon task modeled after Gerardi-Caulton (2000). The *Animal Go No-Go* task (inhibitory control) presents a series of pictures of animals to children and asks that they click a button every time that they see an animal (go trials) except when that animal is a pig (no-go trials). The *Something's the Same* task (attention shifting) presents children with a page containing two pictures that are similar along one dimension (content, color, or size) and asks them to indicate which of these pictures is similar to a third picture. Item response theory models were used to create an overall accuracy score for each task (see Willoughby et al., 2010).

2.4 Analytic Strategy

All study questions were addressed using structural equation models (SEM). An initial confirmatory factor analysis (CFA) model evaluated the fit of the proposed four-factor structure—activity level, fear reactivity, anger reactivity, and regulation using the total sample. Next, a series of multiple-group CFA models were then estimated in order to test the measurement invariance of the 4-factor temperament model separately as a function of child gender, race, and family poverty status. We followed the conventional sequence of models necessary for testing measurement invariance and adopted the parameterization of Reise and colleagues (1993) to facilitate inferences regarding group differences in latent means. Briefly, an initial baseline model involved simultaneously estimating the CFA model across groups without cross-group constraints (i.e., configural invariance). Subsequently, cross-group constraints were imposed on factor loadings (i.e., weak invariance) and then a combination of factor loadings and item intercepts (i.e., strong invariance). Likelihood ratio tests are used to test whether the imposition of cross-group constraints results in a decrement in model fit. Significant nested likelihood ratio tests provide evidence that at least a subset of the constrained parameters differ across groups (e.g., some tasks may be more strongly associated with a latent construct for one versus the other group).

Models in which a subset of indicators differ across groups are referred to as partially invariant (Byrne, Shavelson, & Muthén, 1989). We tested group differences in latent mean levels of all constructs that exhibited (partial) strong invariance (constructs that do not exhibit at least partial strong invariance are not measured on the same metric which renders group comparisons uninterpretable). It is well established that nested likelihood ratio tests are over-powered for samples as large as that used here, which has the potential to result in the appearance of differences in model parameters that are of trivial substantive importance. As such, we utilized $p < .01$ as a threshold for defining significant model. Finally, following the logic of Little (1997), we tested for cross-group equivalence of latent (co)variances, which informed whether some aspects of latent temperament were more variable or differentially associated (e.g., more or less correlated) across groups, and latent means, which provided an estimate of group differences that were not attenuated by measurement error. All SEMs utilized a robust full information maximum likelihood estimator and took into account the complex sampling design (stratification and individual probability weights). Model fit was evaluated using the likelihood ratio chi square test, as well as Comparative Fit Index (CFI) and the root mean squared error of approximation (RMSEA) fit indices, where values of CFI ≥ .95 or RMSEA ≤ .05 were indicative of good fit (Hu & Bentler, 1999; Yu, 2003). All models were estimated using version 7.1 of *Mplus* software (Muthén & Muthén, 1998–2013).

3.0 Results

3.1 Descriptive Statistics

A total of 22 indicators were used to represent individual differences in activity level (4 indicators; home visitor ratings at 6, 15, 24 months and observed activity level from parent-child interaction at 6 months), fear reactivity (4 indicators; home visitor ratings and observed reactivity from the mask task at 15 and 24 months), anger reactivity (4 indicators;

home visitor rating and observed reactivity from the toy removal task at 15 and 24 months), and regulation (10 indicators; one home visitor rating, two direct assessments, and one observation of persistence at 24 months, plus one home visitor rating, four direct assessments, and one observation of persistence at 36 months). Descriptive statistics for each indicator are summarized in Table 1. Consistent with our use of multiple informants and methods, as well as items that spanned assessment occasions, bivariate correlations were of weak to moderate magnitude among indicators of each construct (|*r*|s ≈ .00 – .30), as well as among indicators across constructs (|*r*|s ≈ .00 – .20).

3.2 Confirmatory Factor Analysis: Total Sample

A 4-factor CFA model resulted in acceptable model fit, χ^2 (191) = 446.10; CFI = .91, RMSEA = .03. None of the model implied modifications indices were conceptually defensible. All of the variances for latent variables were statistically significant (ps < .05), which indicated individual differences for all four constructs. As summarized in Table 2, with the exception of one indicator for fear reactivity (RA ratings of fear at the 15 month visit), all of the proposed factor loadings were statistically significant (ps < .05). Consistent with the weak to modest bivariate correlations among indicators that were observed above, most of the standard factor loadings were of modest to moderate magnitude (standardized λs ≈ .20 – .60). Each construct was defined by indicators that were drawn from multiple assessment periods (6–24 months for activity level; 15–24 months for fear and anger reactivity; 24–36 months for regulation); hence, temperamental constructs represented stable individual differences in reactivity and regulation that were observed across the first three years of life. Moreover, with the exception of fear reactivity, each construct was defined by a combination of RA ratings and children's observed performance on challenge or regulatory tasks. Fear reactivity was defined nearly exclusively from children's observed reactions to the mask task at the 15 and 24 month assessments; RA ratings made negligible contributions.

The bivariate associations between latent variables of activity level, anger and fear reactivity, and regulation were of moderate magnitude (see Table 3).

Among the three dimensions of reactivity, activity level was uncorrelated with fear (φ = −.07, p = .25) and only weakly positively correlated with anger (φ = .17, p = .02). In contrast, fear and anger reactivity were moderately positively correlated (φ = .44, p <.001). Moreover, whereas activity level was weakly, positively correlated with regulation (φ = .15, p = .04), fear was modestly negatively correlated with regulation (φ = −.19, p = .001) and anger was moderately negatively correlated with regulation (φ = −.54, p <.001.). Item residuals were correlated when they relied on the same reporter within the same time point. Thus, observer reports of regulation, anger, activity, and fear were correlated at 15 months and at 24 months, over and above the correlation implied by the latent constructs

3.3 Confirmatory Factor Analysis: Measurement Invariance

A series of multiple groups CFA models were estimated in order to test the assumption that our set of 22 measures were equally good indicators of the latent constructs of activity level, fear and anger reactivity, and regulation as a function of child gender, race, and household poverty status. By "equally good" we mean that the indicators did not exhibit any systematic biases in level or degree of association with the latent construct.

3.31 Gender—An initial multiple-group CFA model that was simultaneously estimated for boys (N = 616) and girls (N = 589) provided an adequate fit to the data, χ^2 (384) = 656.01; CFI = .91, RMSEA = .03 (configural invariance). As is summarized in Table 4, the factor loadings for all 22 indicators could be equated without degrading model fit (weak invariance); moreover, the item intercepts could equated for 16 of the indicators without appreciably degrading model fit (partial strong invariance). Collectively, these results indicated that the four-factor model of temperamental reactivity was measured equivalently for boys and girls.

3.32 Poverty—An initial multiple-group CFA model that was simultaneously estimated for children who resided in poor (N = 941) and not poor (N = 264) homes provided an adequate fit to the data, χ^2 (384) = 715.30; CFI = .90, RMSEA = .04 (configural invariance). As is summarized in Table 4, the factor loadings and item intercepts for all 22 indicators could be equated without degrading model fit (weak and strong invariance, respectively). Collectively, these results indicated that the four-factor model of temperamental reactivity was measured equivalently for children from low income and not low income homes.

3.33 Race—An initial multiple-group CFA model that was simultaneously estimated for black (N = 521) and white (N = 684) children provided an adequate fit to the data, χ^2(384) = 727.73; CFI = .90, RMSEA = .04 (configural invariance). As is summarized in Table 4, the factor loadings for all 24 indicators could be equated without degrading model fit (weak invariance); moreover, the item intercepts could equated for 11 of the indicators without appreciably degrading model fit (partial strong invariance). Collectively, these results indicated that the four-factor model of temperamental reactivity was measured equivalently for children from different racial backgrounds.

3.4 Confirmatory Factor Analysis: Group Differences in Latent Means and (Co)Variances

Having established comparable measurement of the latent constructs across groups, we next investigated group differences in the degree of association between latent constructs, as well as group differences in latent mean levels of temperamental reactivity and regulation. As summarized in Table 4, the latent variances and covariances could be constrained to equality for all two-group comparisons without a significant decrement to fit (using $p < .01$ as the criterion). Group differences in latent means are depicted in Figure 1. While all of the latent variables in the reference group (i.e., female; white; not poor) were fixed to have a mean equal to zero and a variance equal to one (following the parameterization by Reise et al. 2003), they were freely estimated in the comparison group (i.e., male; black; poor). As such, the estimated latent means in the comparison group were interpretable in standard deviation units of the reference group; this is equivalent to a Cohen's d effect size metric. Males were observed to exhibit significantly higher levels of anger reactivity ($\mu = .37$, $p<.001$) and activity ($\mu = .23$, $p = .02$) and significantly lower levels of regulation ($\mu = -.64$, $p<.001$) than females. Black children were observed to exhibit significantly higher levels of fear reactivity ($\mu =.26$, $p=.001$) and significantly lower levels of regulation ($\mu = -.97$, $p<.001$) than white children. Children who resided in low income households were observed to exhibit significantly higher levels of anger reactivity ($\mu = .32$, $p = .001$) and significantly lower levels of activity ($\mu = -.42$, $p<.001$) and regulation ($\mu = -.80$, $p<.001$) relative to children who did not reside in low income homes.

Because race and poverty status were partially confounded in this sample (the poorest families were disproportionally African American), latent mean differences between racial groups were ambiguous. As a final step, we estimated a structural equation model in which latent constructs were simultaneously regressed on child sex, race, and poverty status in order to delineate the relative contributions of each factor. We accounted for measurement non-invariance of item intercepts using the approach described by Bauer and Hussong (2009). The results of the structural model revealed that the previously significant race differences were substantially reduced in magnitude once poverty was included as a simultaneous predictor (i.e., standardized |β|s .14 for all race comparisons; see Figure 2).

Collectively, demographic factors explained more variation in regulation ($R^2 = .25$) than in specific dimensions of reactivity ($R^2 = .02 – .06$).

4.0 Discussion

This study tested the factor structure and measurement invariance of in children’s temperament using observational indicators that were obtained across the first three years of life. Results were generally consistent with a larger literature that has addressed these same questions primarily using parent report methods. Specifically, observational indicators of temperament were reasonably represented by a multi-factorial model that distinguished three aspects of reactivity (activity, anger, fear) and a general indicator of regulation. The measurement characteristics model generally fit equally well for children as a function of gender, race, and household income level, and group differences were evident for gender and household income level. Each of these results is elaborated in turn.

A guiding objective of model development was to rely on multi-method, multi-informant indicators, which were drawn from two or more assessment occasions (6–24 months for reactivity, 24–36 months for regulation), in order to represent individual differences in activity level, fear and anger reactivity, and regulation. This approach is consistent with the expectation that individual differences in temperament should be evident early in life and should be observable across informants/methods and time. All four latent constructs were defined by observable behaviors that were derived from two (and in the case of activity level three) assessment periods. Moreover, three of the four latent constructs were derived from multi-informant/method data; the exception was fear reactivity, which was defined nearly exclusively by observed responses to the mask tasks at 15 and 24 months (home visitor ratings made negligible contributions). In general, standardized factor loadings were of low to moderate magnitude (see Table 2), which reflected the modest correlations among the indicators for each construct. Our use of multiple indicators per construct, which spanned assessment occasion, represented one strategy for offsetting the weak associations between indicators. Despite the fact that parent report and observational indicators are typically only modestly correlated, the results of this study add to a growing body of evidence that the structure of temperament is comparable across measurement methods.

Whereas fear and anger reactivity were moderately positively correlated with each other, both were weakly associated with activity level. The positive correlation between fear and anger reactivity may reflect individual differences in general dysregulation and may be a “halo” effect in which the rater generalizes across characteristics. For example, parents who observe their infants to crying to novel stimuli may rate their infants similarly as to when their infants cry because they are angry (e.g., when placed in a car seat). Typically, parents’ reports of their children’s anger and fear reactivity are significantly related (e.g., Garstein & Rothbart, 2003) whereas laboratory assessments of these forms of negative reactivity are weakly or unrelated with each other (Gagne, Hull, Aksan, Essex, & Goldsmith, 2011). Our use of home-visitor ratings may have contributed to the positive correlation between anger and fear reactivity.

Regulation was positively related to activity level, whereas it was negatively related to fear and anger reactivity, although the relationship between fear and regulation was weak. The weak negative association between fear reactivity and regulation was surprising given that fear is considered a passive form of regulation that is related to more active, effortful regulation ((Kochanska & Knaack, 2003; Rothbart & Bates, 1998; Stifter & Spinrad, 2002). However, the interpretation that high levels of reactivity influences the development of regulation cannot be made without examining transactional and longitudinal processes.

Tests of measurement invariance largely supported the expectation that these variables were equally good indicators of latent constructs for male and female, black and white, and low and not-low income children. When differences in the measurement characteristics were evident across groups, they tended to be of small magnitude and were likely detectable given the large sample size (and hence excellent statistical power). Establishing that all temperamental latent constructs exhibited at least partial measurement invariance as a function of gender, race, and poverty increased our confidence in the generalizability of using the reactivity and regulation distinction among children from diverse backgrounds.

Relative to female children, male children exhibited significantly higher levels of anger reactivity and activity level and lower levels of regulation (see Figure 1). Differences in activity level and regulation were consistent with the meta-analytic results of Else-Quest et al., (2006) and adds to the most current research examining gender differences across parent ratings, observer ratings, and observational measures (Gagne, Miller & Goldmsith, 2013; Oliuo, Durbin, Klein, Hayden, Dyson, 2013). In contrast, gender differences in anger reactivity found in the present study were not evidenced in the Else-Quest meta-analysis of parent reports but were consistent with findings from Olino and colleagues (Olino, et al., 2013) who found boys demonstrating greater anger and sadness reactivity in the laboratory. Early emerging sex differences in anger reactivity, activity level, and regulation may help to explain the disproportional risk for subsequent disruptive behaviors disorders (Eme, 2009; Eme, 1979, 2007), including perhaps ADHD (Ramtekkar, Reiersen, Todorov, & Todd, 2010; Rucklidge, 2010), among males.

Previous studies that investigated temperament differences as a function of socioeconomic status yielded mixed results, though a recurring idea has been that children from disadvantaged backgrounds exhibited more “difficult” temperament. Importantly, all of these comparisons were limited to parent report data. In this study, children from low-income households exhibited higher levels of anger reactivity, lower levels of activity, and lower levels of regulation. These results are consistent with a growing literature that has documented that poverty negatively impacts emerging regulatory capacity in young children (e.g., Evans & Kim, 2013; Kishiyama, Boyce, Jimenez, Perry, & Knight, 2009; Li-Grining, 2007). Children from low income families are disproportionately more likely to exhibit conduct problems in childhood, and this association may be causal (Costello, Compton, Keeler, & Angold, 2003; Costello, Erkanli, Copeland, & Angold, 2010; D’Onofrio et al., 2009). The results of the current study raise questions about whether the effects of poverty on higher anger reactivity and lower regulation may contribute to the emergence of conduct problems among low income children.

Although we did not have hypotheses about race differences in temperament, we considered race differences given the prominence that race played in the sampling plan. African American children exhibited slightly higher levels of fear reactivity and large differences in regulation. Given the confounding of race and income in this study (the poorest families in this sample were disproportionately African American), we also estimated a structural equation model in which the four dimensions of temperament were simultaneously regressed on child sex, race, and poverty status. These models clarified that after taking poverty status into account, the previously reported race differences in fear reactivity and regulation were markedly reduced in magnitude (see Figure 2). Although controlling for a dichotomous indicator of poverty reduced the magnitude of race differences, we speculate that more nuanced considerations of family resources would account for the small remaining differences in temperament related to race (e.g., Kainz, Willoughby, Vernon-Feagans, & Burchinal, 2012).

Collectively, demographic comparisons were associated with small to modest sized differences in temperamental reactivity but large sized differences in regulation. Understanding the developmental processes that help account for these differences is an important direction for future research, particularly as they relate to maladjustment. The role of temperament in the development of psychopathology has a rich tradition but gaps remain (Stifter & Dollar, in press). For example, the majority of studies have used smaller (compared to the present study) community samples leaving open the question as to whether temperament is a risk factor for mental health disorders. Understanding that certain demographic characteristics and temperament characteristics are both related to childhood behavior problems as well as to each other may inform preventative interventions by narrowing the sample that might most benefit from such programs.

4.1 Limitations

This study is characterized by at least five limitations. First, whereas each of the dimensions of temperamental reactivity was measured using 4 indicators, regulation was measured using 10 indicators. Given the relatively modest correlations among indicators, the regulation construct was likely measured with better precision than reactivity constructs. Second, 6 of the indicators of regulation (i.e., direct assessments of inhibitory control and attention shifting—but not working memory—at 24 and 36 months) have previously been characterized as executive function tasks. We utilized these tasks here because of their strong conceptual overlap with measures of effortful control that have been used elsewhere (e.g., Murray & Kochanska, 2002) and because we wanted to err on the use of more versus fewer indicators of each construct. Although we agree with others that the constructs of executive function and effortful control are more similar than different (Zhou, Chen, & Main, 2012), we acknowledge the lack of conceptual clarity in our use of these tasks in this study. The absence of inhibitory control tasks that involved an affective component (e.g., delay of gratification tasks) at the age 3 (as was available at the age 2 visit) was an especially important omission. Third, some of the ratings of temperament were based on single items that were averaged across home visitor and visits (i.e., two home visits occurred when children were 6, 24, and 36 months old). Multi-item scales would have yielded improved measurement. Fourth, although all four latent constructs included indicators that spanned two or more assessment occasions, we did not explicitly consider developmental changes in temperament across time. Understanding the developmental processes that contribute to stability and change in reactivity and regulation are important areas for future research. Fifth, the FLP is a representative sample of young children drawn from non-metropolitan, low-resource counties. It is not clear to what extent these findings may generalize to children from other settings.

4.2 Conclusions

In sum, a large body of research supports the meta-organization of temperament into dimensions of reactivity and regulation. Whereas most of the evidence in support of this distinction relied exclusively on parent reports of temperament, the current study adds to a growing number of studies that have demonstrated that the same organizational structure is evident from studies that exclusively utilized observational indicators of temperament. Despite important differences in parent-report versus observational methods, the general consistency in the organization of temperamental constructs across studies using different methods is noteworthy. In addition to relying exclusive on observational indicators of temperament, the current study was unique in the utilization of multiple indicators per construct that were drawn from repeated assessments across the first three years of life in a representative sample—characteristics that distinguish it from other birth cohort studies (Thompson et al., 2010). We demonstrated that the structure of temperamental reactivity and regulation were largely comparable for male and female, black and white, and low and not-low income children. Small to moderate sized group differences were evident for reactivity, while moderate to large sized group differences were evident for regulation. Due to our use of latent variable analyses, these effects were not attenuated by measurement error. Future analyses in this sample will consider how individual differences in temperament predict later dimensions of behavioral development, including which experiences may moderate these associations.

Acknowledgments

Support for this research was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P01-HD039667, R03-HD71646), with co-funding from the National Institute on Drug Abuse. The Family Life Project Key Investigators in the first phase of the study included Lynne Vernon-Feagans, Martha Cox, Clancy Blair, Peg Burchinal, Linda Burton, Keith Crnic, Ann Crouter, Patricia Garrett-Peters, Mark Greenberg, Stephanie Lanza, Roger Mills-Koonce, Debra Skinner, Cynthia Stifter, Emily Werner, and Michael Willoughby.

References

Bauer DJ, Hussong AM. Psychometric Approaches for Developing Commensurate Measures Across Independent Studies: Traditional and New Models. Psychological Methods. 2009; 14(2):101–125.10.1037/A0015583 [PubMed: 19485624]

Bell MA, Wolfe CD. Emotion and cognition: An intricately bound developmental process. Child Development. 2004; 75:366–370. [PubMed: 15056192]

Bijttebier P, Roeyers H. Temperament and Vulnerability to Psychopathology: Introduction to the Special Section. Journal of Abnormal Child Psychology. 2009; 37(3):305–308.10.1007/s10802-009-9308-2 [PubMed: 19225877]

Blair MM, Glynn LM, Sandman CA, Davis EP. Prenatal maternal anxiety and early childhood temperament. Stress-the International Journal on the Biology of Stress. 2011; 14(6):644–651.10.3109/10253890.2011.594121

Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychological Bulletin. 1989; 105(3):456–466.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 1959; 5:81–105. [PubMed: 13634291]

Campbell DW, Eaton WO. Sex differences in the activity level of infants. Infant and Child Development. 1999; 8(1):1–17.10.1002/(Sici)1522-7219(199903)8:1<1::Aid-Icd186>3.0.Co;2-O

Carlson SM, Mandell DJ, Williams L. Executive function and theory of mind: Stability and prediction from ages 2 to 3. Developmental Psychology. 2004; 40(6):1105–1122. [PubMed: 15535760]

Carnicero J, Perez-Lopez J, Del Carmen M, Salinas M, Martinez-Fuentes M. A longitudinal study of temperament in infancy: Stability and convergence of measures. European Journal of Personality. 2000; 14:21–37.

Carter AS, Briggs-Gowan MJ, Jones SM, Little TD. The Infant-Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. Journal of Abnormal Child Psychology. 2003; 31(5):495–514.10.1023/A:1025449031360 [PubMed: 14561058]

Costello EJ, Compton SN, Keeler G, Angold A. Relationships between poverty and psychopathology - A natural experiment. Jama-Journal of the American Medical Association. 2003; 290(15):2023–2029.10.1001/jama.290.15.2023

Costello EJ, Erkanli A, Copeland W, Angold A. Association of Family Income Supplements in Adolescence With Development of Psychiatric and Substance Use Disorders in Adulthood Among an American Indian Population. Jama-Journal of the American Medical Association. 2010; 303(19):1954–1960.

D'Onofrio BM, Goodnight JA, Van Hulle CA, Rodgers JL, Rathouz PJ, Waldman ID, Lahey BB. A Quasi-Experimental Analysis of the Association Between Family Income and Offspring Conduct Problems. Journal of Abnormal Child Psychology. 2009; 37(3):415–429.10.1007/s10802-008-9280-2 [PubMed: 19023655]

Durbin CE, Hayden EP, Klein DN, Olino TM. Stability of laboratory-assessed temperamental emotionality traits from ages 3 to 7. Emotion. 2007; 7(2):388–399.10.1037/1528-3542.7.2.388 [PubMed: 17516816]

Dyson MW, Olino TM, Durbin CE, Goldsmith HH, Klein DN. The Structure of Temperament in Preschoolers: A Two-Stage Factor Analytic Approach. Emotion. 2012; 12(1):44–57.10.1037/A0025023 [PubMed: 21859196]

Else-Quest NM, Hyde JS, Goldsmith HH, Van Hulle CA. Gender differences in temperament: A meta-analysis. Psychological Bulletin. 2006; 132(1):33–72. [PubMed: 16435957]

Eme R. Male life-course persistent antisocial behavior: A review of neurodevelopmental factors. Aggression and Violent Behavior. 2009; 14(5):348–358.10.1016/j.avb.2009.06.003

Eme RF. Sex-Differences in Childhood Psychopathology - a Review. Psychological Bulletin. 1979; 86(3):574–595.10.1037/0033-2909.86.3.574 [PubMed: 377358]

Eme RF. Sex differences in child-onset, life-course-persistent conduct disorder. A review of biological influences. Clinical Psychology Review. 2007; 27(5):607–627.10.1016/j.cpr.2007.02.001 [PubMed: 17331630]

Evans GW, Kim P. Childhood Poverty, Chronic Stress, Self-Regulation, and Coping. Child Development Perspectives. 2013; 7(1):43–48.10.1111/Cdep.12013

Forman D, O'Hara M, Larsen K, Coy K, Gorman L, Stuart S. Infant emotionality: Observational methods and the validty of maternal reports. Infancy. 2003; 4:541–565.

Gagne JR, Van Hulle CA, Aksan N, Essex MJ, Goldsmith HH. Deriving Childhood Temperament Measures From Emotion-Eliciting Behavioral Episodes: Scale Construction and Initial Validation. Psychological Assessment. 2011; 23(2):337–353.10.1037/A0021746 [PubMed: 21480723]

Garstein, M.; Bridgett, D.; Low, C. Asking questions about temperament: Self-and other-report measures across the lifespan. In: Zetner, M.; Shiner, R., editors. Handbook of temperament. New York: Guilford; New York: Guilford; 2012. p. 183-207.p. 183-207.

Gerardi Caulton G. Sensitivity to spatial conflict and the development of self-regulation in children 24–36 months of age. Developmental Science. 2000; 3(4):397–404.

Goldsmith, H.; Rothbart, M. The Laboratory Temperament Assessment Battery: Prelocomotor Version. 3.0. Madison, WI: 1996.

Goldsmith HH, Buss AH, Plomin R, Rothbart MK, Thomas A, Chess S, McCall RB. Roundtable: What is temperament? Four approaches. Child Development. 1987; 58:505–529. [PubMed: 3829791]

Grey KR, Davis EP, Sandman CA, Glynn LM. Human milk cortisol is associated with infant temperament. Psychoneuroendocrinology. 2013; 38(7):1178–1185.10.1016/j.psyneuen.2012.11.002 [PubMed: 23265309]

Hackman DA, Farah MJ. Socioeconomic status and the developing brain. Trends in Cognitive Sciences. 2009; 13(2):65–73.10.1016/j.tics.2008.11.003 [PubMed: 19135405]

Henderson HA, Wachs TD. Temperament theory and the study of cognition-emotion interactions across development. Developmental Review. 2007; 27(3):396–427.10.1016/J.Dr.2007.06.004

Henrichs J, Schenk JJ, Schmidt HG, Velders FP, Hofman A, Jaddoe VWV, Tiemeier H. Maternal Pre- and Postnatal Anxiety and Infant Temperament. The Generation R Study. Infant and Child Development. 2009; 18(6):556–572.10.1002/Icd.639

Hu, L-t; Bentler, PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional versus new alternatives. Structural Equation Modeling. 1999; 6(1):1–55.

Jansen PW, Raat H, Mackenbach JP, Jaddoe VWV, Hofman A, Verhulst FC, Tiemeier H. Socioeconomic inequalities in infant temperament The Generation R Study. Social Psychiatry and Psychiatric Epidemiology. 2009; 44(2):87–95.10.1007/s00127-008-0416-z [PubMed: 18663396]

Kainz K, Willoughby MT, Vernon-Feagans L, Burchinal MR. Modeling family economic conditions and young children's development in rural United States: Implications for poverty research. Journal of Family and Economic Issues. 2012; 33:410–420.

Kaplan BJ, Giesbrecht GF, Leung BMY, Field CJ, Dewey D, Bell RC, Team AS. The Alberta Pregnancy Outcomes and Nutrition (APrON) cohort study: rationale and methods. Maternal and Child Nutrition. 2014; 10(1):44–60.10.1111/j.1740-8709.2012.00433.x [PubMed: 22805165]

Kim S, Brody GH, Murry VM. Factor structure of the early adolescent temperament questionnaire and measurement invariance across gender. Journal of Early Adolescence. 2003; 23(3):268–294.10.1177/0272431603254178

Kishiyama MM, Boyce WT, Jimenez AM, Perry LM, Knight RT. Socioeconomic Disparities Affect Prefrontal Function in Children. Journal of Cognitive Neuroscience. 2009; 21(6):1106–1115.10.1162/jocn.2009.21101 [PubMed: 18752394]

Kochanska G, Knaack A. Effortful control as a personality characteristic of young children: Antecedents, correlates, and consequences. Journal of Personality. 2003; 71(6):1087–1112. [PubMed: 14633059]

Kochanska G, Murray KT, Harlan ET. Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. Developmental Psychology. 2000; 36(2): 220–232. [PubMed: 10749079]

Kotelnikova Y, Olino TM, Mackrell SVM, Jordan PL, Hayden EP. Structure of observed temperament in middle childhood. Journal of Research in Personality. 2013; 47(5):524–532.10.1016/j.jrp. 2013.04.013

Leerkes E, Crockenberg S. The impact of maternal characteristics and sensitivity on the concordance between maternal reports and laboratory observations of infant negative emotionality. Infancy. 2003; 4:517–539.

Li-Grining CP. Effortful control among low-income preschoolers in three cities: Stability, change, and individual differences. Developmental Psychology. 2007; 43(1):208–221. [PubMed: 17201520]

Little TD. Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research. 1997; 32(1):53–76.

Martel MM. Research Review: A new perspective on attention-deficit/hyperactivity disorder: emotion dysregulation and trait models. Journal of Child Psychology and Psychiatry. 2009; 50(9):1042–1051. [PubMed: 19508495]

Matheny AP. A longitudinal twin study of stability of components from Bayley's Infant Behavior Record. Child Development. 1983; 54:356–360. [PubMed: 6683619]

Mayes L. A behavioral teratogenic model of the impact of prenatal cocaine exposure on arousal regulatory systems. Neurotoxicology and Teratology. 2002; 24:385–395. [PubMed: 12009493]

Maziade M, Boudreault M, Thivierge J, Caperaa P, Cote R. Infant Temperament - Ses and Gender Differences and Reliability of Measurement in a Large Quebec Sample. Merrill-Palmer Quarterly-Journal of Developmental Psychology. 1984; 30(2):213–226.

Mebert CJ. Dimensions of Subjectivity in Parents Ratings of Infant Temperament. Child Development. 1991; 62(2):352–361. [PubMed: 2055126]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrica. 1993; 58(4):525–543.

Muris P, Ollendick TH. The role of temperament in the etiology of child psychopathology. Clinical Child and Family Psychology Review. 2005; 8(4):271–289.10.1007/s10567-005-8809-y [PubMed: 16362256]

Murray KT, Kochanska G. Effortful control: Factor structure and relation to externalizing and internalizing behaviors. Journal of Abnormal Child Psychology. 2002; 30(5):503–514. [PubMed: 12403153]

Muthén, LK.; Muthén, BO. Mplus Users Guide. 7. Los Angeles, CA: 1998–2013.

Nigg JT. Temperament and developmental psychopathology. Journal of Child Psychology and Psychiatry. 2006; 47(3–4):395–422. [PubMed: 16492265]

Oakland T, Lu L. Temperament styles of children from the people's Republic of China and the United States. School Psychology International. 2006; 27(2):192–208.10.1177/0143034306064545

Oberklaid F, Prior M, Sanson A, Sewell J, Kyrios M. Assessment of temperament in the toddler age group. Pediatrics. 1990; 85(4):559–566. [PubMed: 2314969]

Parade SH, Leerkes EM. The reliability and validity of the Infant Behavior Questionnaire-Revised. Infant Behavior & Development. 2008; 31(4):637–646.10.1016/j.infbeh.2008.07.009 [PubMed: 18804873]

Persson-Blennow I, Mcneil TF. Temperament Characteristics of Children in Relation to Gender, Birth-Order, and Social-Class. American Journal of Orthopsychiatry. 1981; 51(4):710–714. [PubMed: 7294175]

Podsakoff PM, MacKenzie SB, Podsakoff NP. Sources of method bias in social science research and recommendations on how to control it. Annual Review of Psychology. 2012; 63:539–569.10.1146/annurev-psych-120710-100452

Putnam SP, Stifter CA. Reactivity and regulation: The impact of Mary Rothbart on the study of temperament. Infant and Child Development. 2008; 17(4):311–320.

Ramtekkar UP, Reiersen AM, Todorov AA, Todd RD. Sex and Age Differences in Attention-Deficit/Hyperactivity Disorder Symptoms and Diagnoses: Implications for DSM-V and ICD-11. Journal of the American Academy of Child and Adolescent Psychiatry. 2010; 49(3):217–228. [PubMed: 20410711]

Rapee RM, Coplan RJ. Conceptual Relations Between Anxiety Disorder and Fearful Temperament. Social Anxiety in Childhood: Bridging Developmental and Clinical Perspectives. 2010; 127:17–31.10.1002/Cd.260

Raver CC. Low-Income Children's Self-Regulation in the Classroom: Scientific Inquiry for Social Change. American Psychologist. 2012; 67(8):681–689. [PubMed: 23163459]

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin. 1993; 114(3):552–566. [PubMed: 8272470]

Richardson GA, Goldschrmidt L, Willford J. The effects of prenatal cocaine use on infant development. Neurotoxicology and Teratology. 2008; 30(2):96–106.10.1016/j.ntt.2007.12.006 [PubMed: 18243651]

Rothbart, M.; Bates, J. Temperament. In: Damon, W.; Eisenberg, N., editors. Handbook of child psychology: Vol. 3. Social, emotional, and personality development. 5. New York: Wiley; 1998. p. 105-176.

Rothbart, MK.; Derryberry, D.; Posner, MI. A psychobiological approach to the development of temperament. In: Bates, JE.; Wachs, TD., editors. Temperament: individual differences at the interface of biology and behavior. Washington DC: American Psychological Association; 1994. p. 83-116.

Rubin KH, Hemphill SA, Chen XY, Hastings P, Sanson A, Lo Coco A, Cui LY. A cross-cultural study of behavioral inhibition in toddlers: East-West-North-South. International Journal of Behavioral Development. 2006; 30(3):219–226.10.1177/0165025406066723

Rucklidge JJ. Gender Differences in Attention-Deficit/Hyperactivity Disorder. Psychiatric Clinics of North America. 2010; 33(2):357. [PubMed: 20385342]

Sameroff AJ, Seifer R, Elias PK. Socio-Cultural Variability in Infant Temperament Ratings. Child Development. 1982; 53(1):164–173.10.1111/j.1467-8624.1982.tb01304.x [PubMed: 7060419]

Sanson A, Prior M, Garino E, Oberklaid F, Sewell J. The Structure of Infant Temperament - Factor-Analysis of the Revised Infant Temperament Questionnaire. Infant Behavior & Development. 1987; 10(1):97–104.10.1016/0163-6383(87)90009-9

Schuetze P, Molnar DS, Eiden RD. Profiles of reactivity in cocaine-exposed children. Journal of Applied Developmental Psychology. 2012; 33(6):282–293.10.1016/j.appdev.2012.08.002 [PubMed: 23204615]

Seifer R, Sameroff A, Dickstein S, Schiller M, Hayden L. Your own children are special: Clues to the sources of reporting bias in temperament assessments. Infant Behavior and Development. 2004; 27:323–341.

Slobodskaya HR, Gartstein MA, Nakagawa A, Putnam SP. Early Temperament in Japan, the United States, and Russia: Do Cross-Cultural Differences Decrease With Age? Journal of Cross-Cultural Psychology. 2013; 44(3):438–460.10.1177/0022022112453316

Stifter C, Corey J. Vagal regulation and observed social behavior in infancy. Social Development. 2001; 10:189–201.

Stifter CA, Braungart JM. The Regulation of Negative Reactivity in Infancy - Function and Development. Developmental Psychology. 1995; 31(3):448–455.10.1037/0012-1649.31.3.448

Stifter CA, Spinrad TL. The Effect of Excessive Crying on the Development of Emotion Regulation. Infancy. 2002; 3(2):133–152.10.1207/S15327078in0302_2

Stifter CA, Willoughby MT, Towe-Goodman N, Investigators F. Agree or agree to disagree? Assessing the convergence between parents and observers on infant temperament. Infant and Child Development. 2008; 17(4):407–426. [PubMed: 19936035]

Strelau, J. The concepts of arousal and arousability as used in temperament studies. In: Bates, JE.; Wachs, TD., editors. Temperament: individual differences at the interface of biology and behavior. Washington DC: American Psychological Association; 1994. p. 117-142.

Thompson L, Kemp J, Wilson P, Pritchett R, Minnis H, Toms-Whittle L, Gillberg C. What have birth cohort studies asked about genetic, pre- and perinatal exposures and child and adolescent onset mental health outcomes? A systematic review. European Child & Adolescent Psychiatry. 2010; 19(1):1–15.10.1007/s00787-009-0045-4 [PubMed: 19636604]

Vaughn B, Bradley C, Joffe L, Seifer R, Barglow P. Maternal characteristics measured prenatally are predicitive of ratings of temperamental "difficulty" on the Carey Infant Temperament Questionnaire. Development Psychology. 1987; 23:152–161.

Vaughn B, Taraldson B, Crichton L, Egeland B. The assessment of infant temperament: A critique of the Carey Infant Temperament Questionnaire. Infant Behavior and Development. 1981; 4:1–17.

Vernon-Feagans L, Cox M, Investigators FLPK. The Family Life Project: An epidemiological and developmental study of young children living in poor rural communities. Monographs of the Society for Research in Child Development. 2013; 78(5):1–150. [PubMed: 24147448]

Weiss SJ, Jonn-Seed MS, Harris-Muchell C. The contribution of fetal drug exposure to temperament: potential teratogenic effects on neuropsychiatric risk. Journal of Child Psychology and Psychiatry. 2007; 48(8):773–784.10.1111/j.1469-7610.2007.01745.x [PubMed: 17683449]

Wessman J, Schonauer S, Miettunen J, Turunen H, Parviainen P, Seppanen JK, Paunio T. Temperament Clusters in a Normal Population: Implications for Health and Disease. PLoS ONE. 2012; 7(7):ARTN e33088.10.1371/journal.pone.0033088

Willoughby MT, Blair CB, Wirth RJ, Greenberg M, Investigators FLP. The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. Psychological Assessment. 2010; 22(2):306–317. [PubMed: 20528058]

Wolfe CD, Bell MA. Working memory and inhibitory control in early childhood: Contributions from physiology, temperament, and language. Developmental Psychobiology. 2004; 44(1):68–83. [PubMed: 14704991]

Yu, C-Y. Evaluating cutoff criteria of model fit indices for latent variables with binary and continuous outcomes. UCLA; Los Angeles: 2003.

Zalewski M, Lengua LJ, Fisher PA, Trancik A, Bush NR, Meltzoff AN. Poverty and Single Parenting: Relations with Preschoolers' Cortisol and Effortful Control. Infant and Child Development. 2012; 21(5):537–554.10.1002/Icd.1759

Zhou Q, Chen SH, Main A. Commonalities and Differences in the Research on Children's Effortful Control and Executive Function: A Call for an Integrated Model of Self-Regulation. Child Development Perspectives. 2012; 6(2):112–121.10.1111/j.1750-8606.2011.00176.x

Zhou Q, Lengua LJ, Wang Y. The Relations of Temperament Reactivity and Effortful Control to Children's Adjustment Problems in China and the United States. Developmental Psychology. 2009; 45(3):724–739.10.1037/A0013776 [PubMed: 19413428]

Zimprich D, Mascherek A. Measurement invariance and age-related differences of trait anger across the adult lifespan. Personality and Individual Differences. 2012; 52(3):334–339.10.1016/j.paid. 2011.10.030

Highlights

- A four-factor model of observed temperamental reactivity and regulation fit the data well
- The model fit equally well for children of different gender, race, and income levels
- Demographic group differences were more pronounced for regulation than reactivity

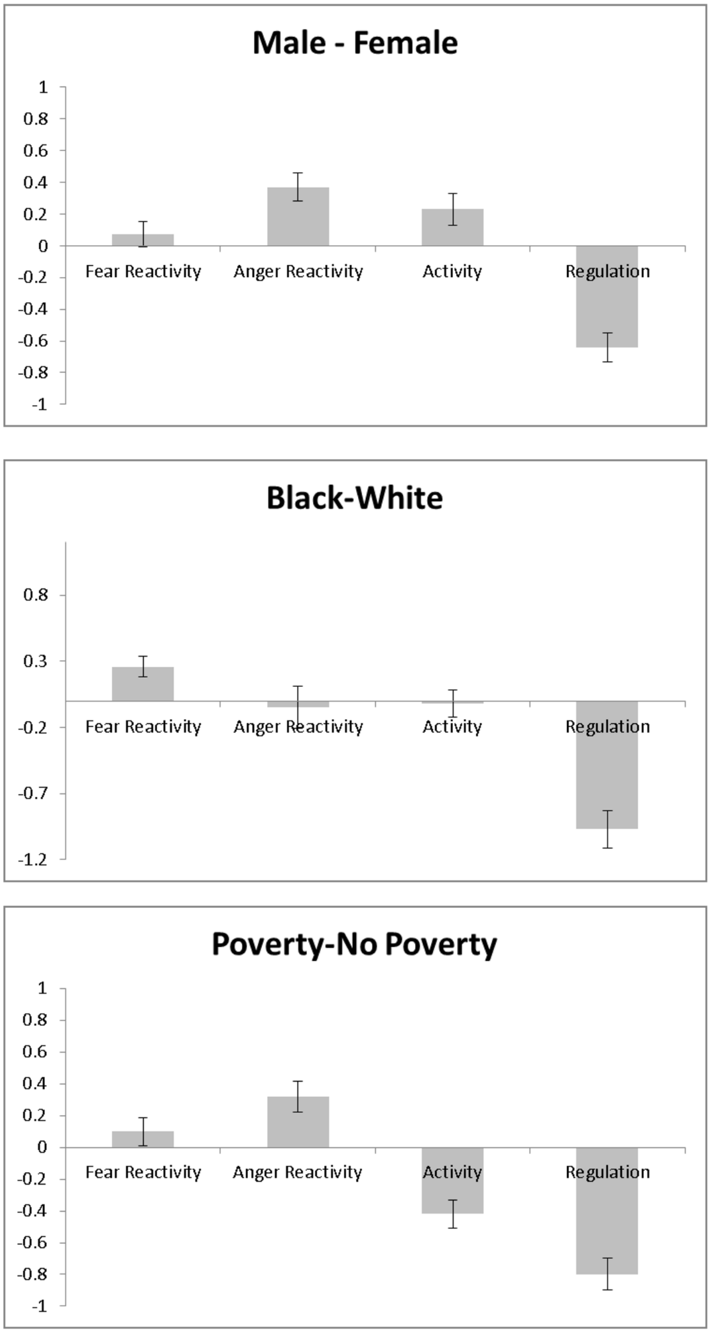

Figure 1.

Latent mean differences in infant temperament by: gender (top), race (middle), and poverty status (bottom). The y-axis reflects the standardized group differences. A positive value indicates that male (top), black (middle), and poor (bottom) children had higher scores, while negative values indicate the reverse.

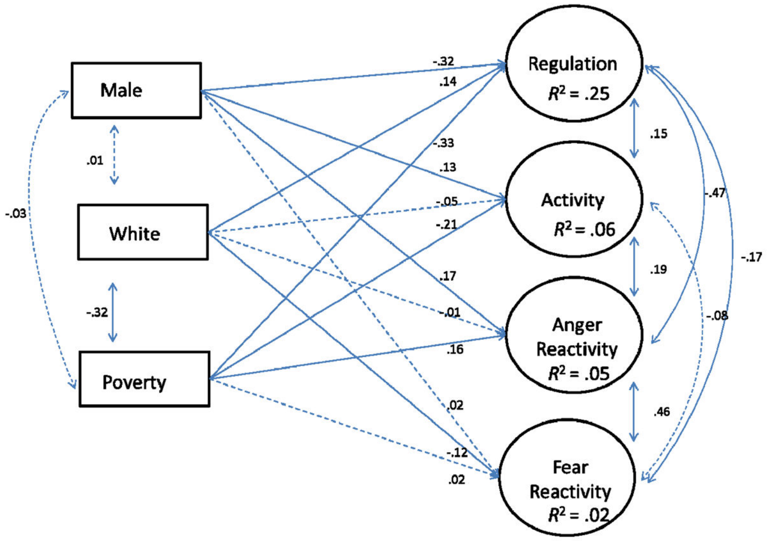

Figure 2.

Standardized structural coefficients in a structural equations model predicting infant temperament from gender, race, and poverty status

Table 1

Descriptive Statistics for Indicator Variables

Construct	Indicator (Month)	N	M	SD	Skew	Kurt
Activity Level	OBS - Activity (6M)	885	2.9	0.9	0.3	−0.3
	RAT - Gross Motor (6M)	933	5.5	1.0	−0.7	1.2
	RAT - Gross Motor (15M)	1083	6.6	0.9	−1.7	5.3
	RAT - Gross Motor (24M)	1069	6.2	1.1	−0.9	1.6
Anger Reactivity	DA - Toy Removal (15M)	959	0.2	0.2	0.9	0.0
	RAT - Irritability (15M)	1083	3.4	1.0	1.1	3.1
	DA - Toy Removal (24M)	868	0.2	0.2	1.7	2.1
	RAT - Irritability (24M)	1069	3.1	1.3	1.2	1.7
Fear Reactivity	DA - Masks (15M)	845	0.2	0.2	0.8	−0.3
	RAT - Reaction to new/strange (15M)	1083	3.1	1.3	1.1	1.9
	DA - Masks (24M)	791	0.3	0.3	0.7	−0.8
	RAT - Reaction to new/strange (24M)	1069	3.2	1.5	0.8	0.2
Regulation	RAT - Persistence (24M)	1069	5.7	1.1	−0.8	0.7
	OBS - Persistence (24M)	1016	3.1	1.1	0.2	−1.1
	DA - Reverse Categorization (24M)	1000	0.8	1.0	0.7	−0.9
	DA - Delay Gratification (24M)	822	1.0	0.7	0.1	−1.1
	RAT - Persistence (36M)	915	5.2	1.5	−0.5	−0.2
	OBS - Persistence (36M)	849	3.3	0.9	−0.1	−0.6
	DA - Silly Sounds Stroop (36M)	378	−0.5	0.8	0.3	−0.9
	DA - Go/No-Go (36M)	350	−0.4	0.98	0.0	−1.1
	DA-Spatial Conflict (36M)	727	−0.0	0.8	−0.4	−0.7
	DA-Something the Same (36M)	687	−0.6	0.8	0.0	−0.7

Note. M = Mean; SD = standard deviation; Kurt = Kurtosis; OBS = Observed; RAT = Home visitor rating; DA = Direct Assessment

Table 2

Standardized Parameter Estimates from Measurement Model in the Total Sample

Construct	Indicator (Month)	Factor Loading	Standard Error
Activity Level	OBS - Activity (6M)	.52***	.05
	RAT - Gross Motor (6M)	.70***	.06
	RAT - Gross Motor (15M)	.34***	.06
	RAT - Gross Motor (24M)	.27***	.05
Anger Reactivity	DA - Toy Removal (15M)	.23***	.07
	RAT - Irritability (15M)	.19***	.06
	DA - Toy Removal (24M)	.46***	.05
	RAT - Irritability (24M)	.61***	.06
Fear Reactivity	DA - Masks (15M)	.39***	.07
	RAT - Reaction to new/strange (15M)	.08	.05
	DA - Masks (24M)	.93***	.15
	RAT - Reaction to new/strange (24M)	.12*	.05
Regulation	RAT - Persistence (24M)	.45***	.04
	OBS - Persistence (24M)	.55***	.04
	DA - Reverse Categorization (24M)	.17***	.04
	DA - Delay Gratification (24M)	.44***	.04
	RAT - Persistence (36M)	.69***	.03
	OBS - Persistence (36M)	.46***	.04
	DA - Silly Sounds Stroop (36M)	.22**	.08
	DA - Go/No-Go (36M)	.19*	.08
	DA-Spatial Conflict (36M)	.45***	.05
	DA-Something the Same (36M)	.45***	.05

Note. N = 1205;

* *p*<.05,

** *p*<.01,

*** *p*<.001;

OBS = Observed; RAT = Home visitor rating; DA = Direct Assessment

Table 3

Latent Correlations between Temperamental Constructs in the Total Sample

	1.	2.	3.	4.
1. Activity Level	--			
2. Fear Reactivity	−.07	--		
3. Anger Reactivity	.17*	.44***	--	
4. Regulation	.15*	−.19***	−.54***	--

Note.

* *p*<.05;

** *p*<.01;

*** *p*<.001

Table 4

Testing for Measurement Invariance and Structural Invariance by Gender, Poverty Status, and Race.

		Model Fit				Model Comparisons		
Group	**Type**	**X^2**	**DF**	**RMSEA**	**CFI**	**LRT**	**DF**	***p*-value**
Gender	Configural	656.01	384	.03	.91			
	Weak	666.56	402	.03	.91	12.14	18	.84
	Strong	708.49	420	.03	.90	40.94	18	<.01
	Partial Strong	691.45	414	.03	.91	24.30	12	.02
	Variances Equal	691.42	418	.03	.91	−.03	4	1.00
	Correlations Equal	699.45	424	.03	.91	8.03	6	.24
Poverty	Configural	715.30	384	.04	.90			
	Weak	736.52	402	.04	.90	23.15	18	.19
	Strong	767.59	420	.04	.89	31.07	18	.03
	Variances Equal	776.88	424	.04	.89	9.29	4	.05
	Correlations Equal	789.53	430	.04	.89	12.65	6	.05
Race	Configural	727.73	384	.04	.90			
	Weak	756.20	402	.04	.89	30.04	18	.04
	Strong	895.72	420	.04	.86	139.52	18	<.01
	Partial Strong	765.84	409	.04	.89	9.64	7	.21
	Variances Equal	775.33	413	.04	.89	8.54	4	.07
	Correlations Equal	782.34	419	.04	.89	7.01	6	.32

Note. Scaling correction has been applied to the LRT due to use of MLR estimator. Each LRT compares the model in the corresponding row to the next-most constrained model that was not significantly worse than its comparison model. Following the parameterization of Reise et al. (1993), 4 indicators (1 per factor) had equated loadings and intercepts in the configural models to facilitate identification; DF=degrees of freedom; RMSEA = Root Mean Squared Error; CFI = Confirmatory Fit Index; LRT = Likelihood Ratio Test