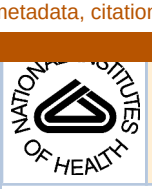

Published in final edited form as:

Infant Child Dev. 2008 August 1; 17(4): 407–426. doi:10.1002/icd.584.

Agree or Agree to Disagree? Assessing the Convergence between Parents and Observers on Infant Temperament

Cynthia A. Stifter[a,*], **Michael T. Willoughby**[b], **Nissa Towe-Goodman**[a], and **The Family Life Project Key Investigators**

[a] The Pennsylvania State University, University Park, PA, USA

[b] The University of North Carolina, Chapel Hill, NC, USA

Abstract

The assessment of infant temperament has been typically accomplished with parent questionnaires. When compared with temperament behaviours observed in the laboratory, parents and observers generally do not agree, leading some researchers to question the validity of parent report. This paper reports on a representative sample of infants whose families resided in non-metropolitan counties and whose temperament was measured in three ways: (1) standard parent report (Infant Behavior Questionnaire); (2) observer ratings across two lengthy home visits; and (3) observer coding of second-by-second reactions to specific emotion-eliciting tasks. In order to account for both trait and method variance, structural equation modelling was applied to a sample of 955 infants (*M* age = 7.3 months) using variables from the three methods that reflected the dimensions of positivity and negativity. Although models based solely on method factors and trait factors fit the data well, results indicated that a model that included method and trait factors provided the best fit. Results also indicated that parents and observers (either across the home visit or to specific tasks) converge, to a degree, on ratings of the positivity dimension but diverge on the negativity dimension.

Keywords

temperament; parent ratings; behavioral observation

It is well acknowledged that Mary Rothbart has made a significant theoretical contribution to the child development field; hence, this current issue of *Infant and Child Development*. Her theory of temperament has paved the way for numerous empirical studies that, in turn, have established the role of individual differences in reactivity and regulation in several developmental domains including attachment, cognition, peer relations, psychopathology, and morality.

Perhaps less recognized, yet highly utilized, are the instruments Rothbart designed to measure temperament across childhood. Beginning with the Infant Behavior Questionnaire (IBQ; Rothbart, 1981), Rothbart designed her questionnaires so as to avoid ‘asking parents either to make global judgments of their child’s behavior or to attempt to recollect occasions of child behavior from the distant past. We did not wish to ask parents to make comparative judgments about their infants…’ (p. 572). Rothbart was able to accomplish this by wording items so that they asked the caretakers to report on the frequency of specific behaviours occurring in the past few weeks. For example, an item would ask the parent to report how frequently during the past week her infant cried in response to being undressed. Such wording was believed to

*Correspondence to: Cynthia A. Stifter, The Pennsylvania State University, University Park, PA, USA. tvr@psu.edu.

increase the accuracy of reporting and reduce response bias. Since the creation of the IBQ, Rothbart has designed questionnaires to tap temperament in children, adolescents, and adults. All of these instruments have demonstrated good to excellent reliability and validity (Gartstein & Rothbart, 2003; Putnam, Gartstein, & Rothbart, 2006; Rothbart, Ahadi, Hershey, & Fisher, 2001; Rothbart, Evans, & Ahadi, 2000). Moreover, the number of studies that have employed these questionnaires to assess temperament is a testament to their utility in the developmental/personality field.

Despite the care with which Rothbart constructed her questionnaires, debate has emerged regarding the validity of parent report of child temperament (Kagan, 1998; Rothbart & Bates, 1998) This is due in large part to the small to nonexistent statistical relations between parent ratings and observed child behaviour. Several studies have reported significant correspondence between parent ratings and laboratory or home observations; however, a similar number have not. Moreover, when significant correlations were found they were generally in the weak to modest range (r's<0.30). These findings have prompted researchers to question whether parents are accurate reporters of their children's temperament and initiate a number of investigations aimed at understanding the source of this lack of concordance (Forman *et al.*, 2003; Seifer, Sameroff, Dickstein, Schiller, & Hayden, 2004; Vaughn, Bradley, Joffe, Seifer, & Barglow, 1987; Vaughn, Taraldson, Crichton, & Egeland, 1981).

There are several reasons why parents and observers may not agree, ranging from social desirability on the part of the parent to limitations on the number of observations that can be conducted in the laboratory. According to Rothbart and Bates (1998) these reasons can be organized around three potential sources of measurement error: (1) rater characteristics that are relatively independent of child behaviour; (2) bias as a function of child behaviour or child–parent interaction; and (3) method factors. Several studies have been conducted to address some of these sources of error. For example, taking a components-of-variance approach, researchers have investigated which factors explain the subjective components of parent ratings (e.g. Bates & Bayles, 1984; Seifer *et al*., 2004; Vaughn *et al*., 1987). Likewise, a handful of studies were designed to increase correspondence between parents and observers (Bohlin, Hagekull, & Lindhagen, 1981; Bornstein, Gaughran, & Sequi, 1991; Seifer, Sameroff, Barrett, & Krafchuk, 1994). The findings, thus far, have been mixed.

A growing body of evidence suggests that parental characteristics may impact the degree of concordance between various measures of infant temperament. Parents' early childhood experiences and depression have been linked to greater discrepancies between parent reports and observational assessments of infant temperament (Forman *et al*., 2003; Leerkes & Crockenberg, 2003), and numerous studies have documented systematic associations between parent personality, psychopathology, and stress with temperament ratings (Mebert, 1991; Sameroff, Seifer, & Elias, 1982; Vaughn *et al*., 1981). Such factors may alter parent perceptions of their children's behaviour, interfering with the ability to accurately identify and report on their actions and emotional responses.

The context within which the parent and the observer are rating the child's behaviour may also explain their lack of agreement. Although parents can observe their children in a variety of situations, the child's behaviour is almost always within the context of the parent–child relationship and the environment that the parent provides, particularly in infancy when the social world of the infant is somewhat constrained. Indeed, several studies have shown that the quality of the parent–child relationship may influence development in such temperament dimensions as positive and negative emotionality (Belsky, Fish, & Isabella, 1991), effortful control (Aksan & Kochanska, 2004), and inhibition (Park, Belsky, Putnam, & Crnic, 1997; Rubin, Burgess, & Hastings, 2002). On the other hand, in the laboratory the novelty of the

situation and the artificiality of the eliciting tasks (e.g. arm restraint, still-face) may influence the child to act in ways that are atypical for that child.

Another reason for the lack of correspondence is that parents view their children across a variety of situations, while laboratory assessments only provide a snapshot of the child's temperament. Thus, even when using a temperament instrument such as the IBQ that constrains questions to recent events, parents have more access to characteristic patterns of responses by the child, while laboratory observations are short-lived, and can be influenced by the state of the infant upon arrival to the lab. Likewise, parents are privy to low-frequency events that cannot, for purposes of time, be captured in the laboratory. Several studies that restricted parents' observations of their children in a particular situation have found improved parent–observer concordance, although not in all dimensions and not to the degree expected (Bohlin *et al*., 1981; Bornstein *et al*., 1991). For example, Seifer *et al*. (1994) had both parents and observers rate infant reactions to the same tasks using a rating system specifically designed for the study. Findings, even when multiple observations were aggregated, continued to suggest that parents and observers do not agree, leading the authors to infer that mothers are 'poor' reporters of their infants' behaviour. This conclusion, however, did not consider differences in methodology. Although parents and observers used the same scale, observers were trained on the rating scale that was applied to videotapes of the sessions, while parents were minimally trained and rated their infants' behaviour at the conclusion of the sessions. Thus, observers had the benefit of previewing and reviewing child behaviour while the parents had to use recall. Likewise, parents typically rate children on a scale with little instruction on the meaning of terms such as 'irritable', 'difficult', and 'active'. On the other hand, observers are trained to reliably identify certain behaviours and often code the frequency or duration with which the behaviours occur (Goldsmith & Hewitt, 2003). Thus, even in cases where the parent and the observer used the same scale (Seifer *et al*., 1994), observers had the added advantage of being trained on the precise meaning of the behaviours they were asked to rate.

Although not an exhaustive list, these three reasons for the lack of concordance between parent report and laboratory observations highlight the measurement problems that confront child temperament researchers. In particular, there is concern that many investigators use parent ratings as measures of temperament without regard to these methodological problems (Seifer, 2002).

In the present study, we attempted to address some of the issues that continue to plague the infant temperament field. Our first aim was to utilize a measure of temperament that was not reliant on parents but mimicked parent ratings in several ways. That is, we searched for a method that was more global in its assessment of temperament and did not require training or the viewing of videotapes of child behavior. Towards that end, we adapted the Infant Behavior Record (IBR; Bayley, 1969). Originally designed to assess infant behaviour during a test of mental development, the IBR has since been recast as a temperament assessment reflecting such dimensions of social orientation, emotional tone, task orientation, and activity level (Goldsmith & Gottesman, 1981; Matheny, 1983) and has been successfully used with twins (Saudino, Plomin, & DeFries, 1996), siblings (Braungart, Plomin, DeFries, & Fulker, 1992), and premature infants (Meisels, Cross, & Plunkett, 1987). Typically, the IBR is rated by the Bayley examiner after the completion of the cognitive assessment. More recently, the IBR has been applied to laboratory settings (Carnicero, Perez-Lopez, Del Carmen, Salinas, & Martinez-Fuentes, 2000; Stifter & Corey, 2001). In the present study, we used the IBR to rate the infants' behaviour across two home visits that lasted approximately 2.5 h each. The purpose of the home visits was to assess infants' reactions to several challenges, as well as collect biological measures on the infant and parent, observe parent–child interactions, and obtain family/individual information from the parents via interviews and questionnaires. Observers were minimally trained on the use of the IBR and completed their ratings at the end of each home

visit. Thus, we obtained an assessment of temperament that included a variety of situations, such as transitions between tasks and during down times when only the mother was engaged with the home visitors.

The second aim of the current study was to examine how well observers, either those using more global methods or those coding specific behaviours from videotape, converged with parent ratings of infant temperament. Previous studies investigating the degree of agreement between parents and observers utilized simple correlations between raters. Correctly or incorrectly, conclusions were drawn that parents were not good reporters of their child's temperament. However, these analyses did not take into account individual differences that are due to trait as well as method factors. In the present study we used structural equation modelling (SEM) to approach this issue. Briefly, we made use of a confirmatory factor model strategy that was developed for the comparison of methods/informants in the context of multitrait-multimethod data (Eid, 2000). This model allowed us to make inferences about how observers' ratings compared with parent reports of temperament, focusing exclusively on the reliable variation in indicators of infant temperament. By reliable variation, we mean variation in temperament indicators that are systematically related to latent traits (positivity, negativity) or general method (informant) effects.

The aims of the present study were accomplished with data from a large, longitudinal study of infants raised in non-metropolitan areas. When infants were approximately 7 months of age they were visited in their homes by two home visitors. Parents completed several subscales of an infant temperament questionnaire at one of the home visits and infants participated in several tasks designed to elicit individual differences in emotional responsiveness, which were videotaped for off-line micro-analytic coding. At the end of each home visit, home visitors rated the infants' temperament using an adapted version of the IBR. We expected that a model that included both trait and method factors would improve the convergence between raters. In addition, we hypothesized that the home visitors, due to a similarity in method, would agree more with parents than coders.

METHOD

Participants

The Family Life Project (FLP) was designed to study families that lived in two of the four major geographical areas of high child rural poverty (Dill, 2001). Specifically, three counties in Eastern North Carolina (NC) and Central Pennsylvania (PA) were selected to be indicative of the Black South and Northern Appalachia, respectively. The FLP adopted a developmental epidemiological design. Complex sampling procedures were used to recruit a representative sample of 1292 families at the time that they gave birth to a child, with low-income families in both states, and African-American families in NC, being over-sampled. Given logistical constraints related to obtaining family income data in the context of hospital screening, family income was dichotomized (low versus not low) for purposes of guiding recruitment. Families were designated as low income if they reported household income <200% poverty rate, use of social services requiring a similar income requirement (e.g. food stamps, Women, Infants and Children (WIC), Medicaid), or had less than a high school education.

One hospital in each county was selected at random (with probability proportional to size when there was more than one hospital in a county). In-person recruitment occurred in all three of the hospitals that delivered babies in the target counties. Phone recruitment occurred for families who resided in target counties but delivered in non-target county hospitals. These families were located through systematic searches of the birth records. At both sites, recruitment occurred 7 days per week over the 12-month recruitment period spanning September 15, 2003–September 14, 2004, using a standardized script and screening protocol.

In total, FLP recruiters identified 5471 (57% NC, 43% PA) women who gave birth to a child during the recruitment period, 72% of which were eligible for the study. Eligibility criteria included primary custodial rights of the target child, residency in target counties, English as the primary language spoken in the home, and no intent to move from the area in the next 3 years. Of those eligible, 68% were willing to be considered for the study. Of those willing to be considered, 58% were invited to participate. Invitations for participation were based on screening information related to income and, in NC, to race. Of those selected to participate, 82% (N = 1292) of families completed their first home visit, at which point they were considered enrolled in the study.

Owing to logistical difficulties, approximately 25% of the infants were not tested at the target age (~6–7 months of age). As the goal of the current study was to examine the concordance among observers of temperament, a construct that changes with development, the sample was further constrained to infants younger than 9 months of age, resulting in a sample of 955 infants. Table 1 reports the sample descriptives including child age, child gender, mother age, mother education, child and mother ethnicity, and income-to-needs ratio.

Procedures

Infants and their parents were visited in their homes twice (within 7–10 days) when the infant was approximately 6–7 months of age. Infants and parents participated in several tasks either together (free play interaction, book reading) or separately (infant: challenge tasks, Bayley exam, health screening, saliva sample; parent: interviews, questionnaire completion, saliva samples). The mothers' ratings of infant temperament, the observer ratings of infant temperament, and the infants' reactions to the challenge tasks are the focus of the present study and are described further.

Challenge tasks—Infants participated in four tasks designed to elicit specific reactions. Two of the tasks, which are the focus of the present study, were drawn from the LAB-TAB (LT; Goldsmith & Rothbart, 1996), presentation of masks and arm restraint. The other two tasks (toy reach and barrier) were not used in the present study, with the exception that home visitors could include infant reactions to these tasks in their IBR assessments.

The *mask task*, which was presented prior to the arm restraint, consisted of the presentation of four unusual masks for 10 s each. During each presentation, the home visitor called to the child using his/her name while moving her head slowly from side to side.

The *arm restraint task* consisted of the home visitor lightly restraining the arms of the child down by the child's side. Infants were placed in a walker (for those infants with low torso control, blankets were placed around the child to keep them upright) and the home visitor restrained the arms from behind the infants. Prior to the start of this task mothers were asked to stay out of the sight of their infants and not verbally interact with them. Infants' arms were restrained up to 2 min or 20 s of hard crying and then released for 1 min. After the arm release minute, mothers were told they could remove their children from the walker and soothe them if needed. All tasks were videotaped for off-line coding.

Measures

Parent ratings of temperament—Selected subscales of the revised version of Rothbart's Infant Behavior Questionnaire (IBQ-R; Gartstein & Rothbart, 2003) were administered to mothers during one of the home visits. Mothers completed the fear/distress to novelty (16 items), distress to limitations (16 items), approach (12 items), duration of orienting (12 items), and falling reactivity/recovery from distress (13 items) subscales. A 7-point Likert scale ranging from never (1) to always (7) was used to rate the frequency with which their child

exhibited the behaviours in the last 2 weeks. The IBQ-R subscales were administered by the home visitor via computer using Blaise software. Item values were averaged to produce a score for each of the subscales. Alphas ranged from 0.48 (falling reactivity) to 0.87 (distress to novelty). For the purposes of the present study only the fear/distress to novelty, distress to limitations, and approach scores were retained.

Observer ratings of temperament—After each home visit, an adaptation of the IBR (Bayley, 1969) was completed independently by both home visitors. In the present study the IBR was applied to behaviour observed globally across the entire home visit (see Stifter & Corey, 2001). The IBR items/scales included in the present study were social approach (three items: responsiveness to persons, to examiner, to caregiver; higher scores indicate greater friendliness and approach-oriented behaviours), positive affect (higher scores indicate more happiness), fear (high scores represent greater fear), and irritability (higher scores indicate greater irritability). As there were two separate home visits for each family at the 6-month assessment and two home visitors involved in each visit, a total of four ratings were made for each child. The resulting mean scores were used as an outcome variable. Cross-rater correlations for the two home visits ranged from 0.51 to 0.65. Cross-visit correlations, averaged across raters, were 0.58 for positivity and 0.42 for negativity.

Observed infant reactivity—Negative reactivity and positive reactivity were coded from the digital images of the mask and arm restraint tasks. Second-by-second coding was accomplished using the Better Coding Approach software (Danville, PA). Three levels of negative reactivity were coded: low, moderate, and high. A composite score for negative reactivity for each task was created by summing the seconds of low, moderate, and high negative reactivity and then calculating the proportion by dividing the sum of all negative reactivity scores by the total time of the task. Along with negative reactivity, the presence/absence of positive reactivity was coded during the masks and arm restraint tasks. The proportion of positivity was calculated for each task by taking the duration of positive reactivity and dividing by the total time of the task. Coders were trained to achieve at least 0.75 (Cohen's κ) reliability on the reactivity coding. Subsequent inter-rater reliability was calculated on 15% of cases using κ coefficients resulting in κ's of 0.86 for the arm restraint task and 0.94 for the masks task.

Analytic Strategy

Our primary questions were answered using SEM methods. SEM models were fit using Mplus version 4.0 (Muthén & Muthén, 2004), which accommodated the complex sampling design (i.e. stratification on income and race; individual probability weights associated with over-sampling of low-income and African-American families). All SEM models were estimated using a robust maximum likelihood estimator (MLR). Missing data were handled using the full information maximum likelihood methods (Arbuckle, 1996). In addition to the significance of the likelihood ratio test statistic, model fit was also evaluated using a combination of absolute (standardized root mean residual, SRMR; root mean-squared error of approximation, RMSEA) and comparative (comparative fit index, CFI) fit indices. Following Hu and Bentler (1999), SRMR values $\leq$0.08, RMSEA values $\leq$0.06, and CFI values $\geq$0.95 were considered supplementary indices of good model fit.

After reviewing descriptive statistics, we present five confirmatory factor analytic (CFA) models. The first model represented temperament data using two substantive factors (positivity and negativity). The second model represented temperament data using three method factors (parent, home-visitor, and observer reports). The third, fourth, and fifth models represented temperament data using a combination of two substantive and two method factors. Each of these latter three models is formally known as the correlated trait correlated method minus one

[CTC(M–1)] model. The CTC(M–1) model was developed for multiple trait multiple method data in which the focus is on a comparison of informants/methods (Eid, 2000; Eid, Lischetzke, & Nussbeck, 2006). In this study, we were primarily interested in the model in which parent report was designated as the comparison method (model 3). However, in order to test whether the IBR was more likely to converge with parent ratings than with observers, we also examined models in which home-visitor and observer reports were designated as the comparison methods (models 4 and 5, respectively).

Briefly, the CTC(M–1) model is structured such that one informant/method is designated as the comparison method (for a complete discussion, see Eid, 2000; Eid *et al*., 2006). Figure 1 depicts this model for the case in which parent report is the comparison method. The trait factors represent the reliable variation in items that are measured by the comparison method (e.g. IBQ positive and negative in Figure 1). That is, the latent variables denoted positive and negative in Figure 1 represent the true score variance of temperament indicators as measured by the comparison method. The method factors represent the reliable variation in items that is due to method-specific influences (e.g. variation unique to home-visitor or observer report in Figure 1). That is, the latent variables denoted as home-visitor and observer reports in Figure 1 represent that part of the variance of an indicator that cannot be predicted by the trait factor (the comparison method) and that is not due to measurement error, but rather to systematic method-specific influences. Together, the trait and method factors represent all of the reliable (true score) variation associated with each item, while the residual variances (not depicted in Figure 1) represent the unreliable variation associated with each item. The residual variances in this case consist not only of measurement error, but also of method effects specific to an individual trait (Eid, Lischetzke, Nussbeck & Trierweiler, 2003).

Important byproducts of the CTC(M–1) model are the resulting consistency and method specificity statistics. Consistency coefficients represent the proportion of reliable variation in an item that is explainable by the comparison method. Method specificity coefficients represent the proportion of reliable variation in an item that cannot be explained by the comparison method and is hence unique to its own method. Consistency and method specificity coefficients sum to 1.0 for each item. Inspection of item reliability, consistency, and method specificity coefficients provides a framework for understanding the convergence/divergence in trait variation as a function of different informants/methods. Another way to think about the CTC (M–1) model is that it is a variance decomposition model. The reliable variation for each item (akin to the R^2 for individual items when they are regressed on latent variables) is partitioned into components specific to trait and method factors.

Variables—The variables we chose to address our goals were determined by three factors: (1) our theoretical approach; (2) the analytic method; and (3) the constraints of the longitudinal study from which the data were drawn. Several emotion/personality researchers have taken a dimensional approach to individual differences in emotional reactions to stimuli (Diener, Smith, & Fujita, 1995; Watson, 1988), such that variations in positive affect occupy one dimension, while variations in negative affect occupy another (cf. Russell & Carroll, 1999). The dimensions of positive and negative reactivity are also hypothesized to underlie temperament (Rothbart & Bates, 1998). Rothbart and others (Belsky, Hsieh, & Crnic, 1996; Goldsmith & Campos, 1990; Putnam & Stifter, 2005) provide developmental evidence that positivity and negativity are separable, independent, and stable dimensions of infant temperament. In the present study, we examined the convergence between parents and observers on the dimensions of positivity and negativity by compositing subscales within the IBQ, items within the IBR, and the negativity and positivity scores from the LT coding. Given that aggregation improves agreement, we expected that parents and observers would show better convergence on broader, over more specific, dimensions. Our second motivation for using the positivity and negativity dimensions came from the analytic method used to address

our study goals. So as not to favour any one method under comparison, we created one indicator per method by trait combination for use in the confirmatory factor models. Although the models that we considered could have accommodated multiple indicators, this would have required a minimum of three indicators per method by trait combination (for example, see Eid *et al*., 2003), which were not available in the current study.

The study from which these data were drawn also constrained our ability to examine specific dimensions of temperament. Because of the burden on parents enrolled in the study, it was decided to select only five subscales from the IBQ-R. These were originally chosen based on conceptual questions; thus, the subscales do not directly map onto the dimensions we chose to analyse. Likewise, the adapted IBR did not directly map onto the parent ratings, in that only fear, but not anger, was rated by home visitors. Lastly, the masks and arm restraint tasks were administered to elicit negative reactivity but positive reactivity was also coded during these tasks. It was not expected that 6–7-month-old infants would react negatively to the masks with the exception of those who were highly sensitive to novelty. Indeed, we observed most infants to show interest and smiling to the masks at this age. Thus, emotional reactivity was not measured in response to more positive stimuli.

For the current analyses, the approach scale from the IBQ was used to represent parent ratings of *IBQ positivity* and the fear/distress to novelty and the distress to limitations subscales were combined to represent parent ratings of *IBQ negativity*. For home-visitor ratings, the social approach and positive affect scores were combined to represent *IBR positivity*, and the fear and irritability scores were combined to represent *IBR negativity*. Finally, observer coding of negative reactivity to both the masks and arm restraint tasks comprised the *LT negativity* variable, while coded positive reactivity during both tasks comprised the *LT positivity* variable.

RESULTS

Descriptive Statistics

A total of six variables were used in our analyses, one each for each method/informant (parent, home visitor, observer) by trait (positive and negative reactivity) combination. Table 2 provides correlations for these six variables. Several significant findings emerged. Most notable was that, of the three methods, the home-visitors' positive and negative reactivity scores were strongly related, while coded negativity and positivity were weakly correlated. Surprisingly, the correlation between parent's ratings of positivity and negativity was near zero. Significant cross-method relations were also revealed, but only between home-visitors' ratings and coded positivity and negativity.

Confirmatory Factor Analyses

Trait-only model—The first CFA model consisted of two trait factors, one each for positivity and negativity. The purpose of this model was to evaluate how well two trait factors, positivity and negativity, represented the observed (co)variation in temperament indicators. The initial model indicated that the residual variances for the two home-visitor items were negative, though not significantly different than 0. This implied that all of their variation was accounted for by the trait factors. As such, the residual variances were fixed to 0 and the model was re-estimated. The likelihood ratio test statistic for this model was significant, indicating model misfit $\chi^2(10) = 34.8$, $p<0.0001$; however, all of the fit indices met conventional criteria for good-fitting models (see Table 3). Inspection of parameter estimates indicated that all of the factor loadings were significant and in the expected direction. Similarly, the latent variances for positivity ($\phi = 0.79$, $p<0.0001$) and negativity ($\phi = 0.78$, $p<0.0001$) were significant, indicating that there was systematic variation in these dimensions. The positive and negative trait factors were significantly negatively correlated ($\phi = -0.62$, $p<0.0001$). Collectively, these

results lend some support to conceptualizing temperament in terms of positive and negative dimensions. However, the significance of the likelihood ratio test statistic implied that trait-only factors did not sufficiently explain all of the observed (co)variation in these temperament data.

Method-only model—The second CFA model consisted of three method factors, one each for home-visitor, parent, and observer reports. The purpose of this model was to evaluate how well three method factors—parent report, home-visitor, and observer reports—represented the observed (co)variation in temperament indicators. As with the trait-only model, the likelihood ratio test statistic was significant, indicating model misfit $\chi^2(6) = 34.6$, $p<0.0001$; however, once again, the fit indices met all conventional criteria for good-fitting models (see Table 3). Inspection of parameter estimates indicated that, with the exception of the parent report of positivity (approach), all of the factor loadings were significant and in the expected direction. Whereas the latent variances for the home-visitor ($\phi = 0.54$, $p<0.0001$) and observer ($\phi = 0.01$, $p = 0.013$) factors were significant, the latent variance for the parent-report factor was not ($\phi = 0.10$, $p = 0.27$). These results indicated that, once item-level variation was partitioned into reliable and unreliable components, there was evidence of individual differences in temperament as measured by home visitors and observers, but not parents. As with the trait-only model, the significance of the likelihood ratio test statistic implied that method-only factors did not sufficiently explain all of the observed (co)variation in these temperament data.

CTC(M–1) model with parent ratings as the comparison method—The third model represented a hybrid of the first two models, in that it included two trait and method factors, with parent report as the comparison method (see Figure 1). This is the CTC(M–1) model that was described above and that is depicted in Figure 1, with the exception that item-level residual variances are omitted solely for purposes of presentation. When first estimated, the residual variances for the parent- and home-visitor-reported negative items were negative, though not significantly different than 0. This implied that all of their variation was accounted for by the combination of trait and method factors. As such, these residual variance parameters were fixed to 0, and the model was re-estimated. This model fit the data well, as indicated by the likelihood ratio test statistic, $\chi^2(5) = 11.5$, $p = 0.04$, and fit indices that were improved relative to those associated with the method-only and trait-only models (see Table 3). The trait-only (model 1) model was formally nested in this hybrid model. A χ^2 difference test indicated that the CTC (M–1) model fit the observed data better than the trait-only model, $\chi^2(5) = 24.47$, $p = 0.0003$.[1]

Inspection of parameter estimates indicated that all of the factor loadings were significant and in the expected direction. Whereas the latent variance for positivity was not significant ($\phi = 0.04$, $p = 0.08$), the latent variance for negativity was significant ($\phi = 0.65$, $p\leq0.0001$). These results indicated that, when only reliable variation in temperament data was considered, there was no evidence for significant individual differences in parent-rated infant positivity. In contrast, there was evidence for significant individual differences in parent-rated infant negativity.

The latent variances for the home-visitor ($\phi = 0.29$, $p<0.0001$) and LT ($\phi = 0.01$, $p = 0.03$) reports were both significant, and these factors were significantly correlated ($\phi = 046$, $p = 00018$). These results indicated that, when only reliable variation in temperament data was considered, there was systematic variation in behaviours associated with home-visitor and observer reports, above and beyond that accounted for by parent report. The positive correlation between these method factors indicate that home-visitor and observer reports tend to be in

[1]Given our use of the MLR estimator, this χ^2 difference test was computed using the adjustments described by Satorra and Bentler (1999).

agreement in terms of their over- or under-reporting of positivity and negativity relative to parent report.

The consistency coefficients (CC) and method specificity (MS) coefficients in the first column of Table 4 describe the degree to which home-visitor- and observer-rated behaviours are explained by parent report (CC) and what is unique to the method (MS). For positive reactivity, the CCs suggest that 27% and 51% of the reliable variation in home-visitor and observer ratings could be explained by parent reports. Inspection of the MSs suggests that the remaining 73% and 49% of reliable variation in home-visitor- and observer-reported positive behaviour was unique to these methods. For negativity, only 1% and 5% of the reliable variation in home-visitor and observer ratings could be explained by parent-reported negative behaviour. The remaining 99% and 95% of the reliable variation in home-visitor- and observer-reported negative behaviour was unique to these methods. Collectively, these results suggest that, whereas parent report of *positivity* is moderately to strongly associated with home-visitor and observer reports, parent reports of *negativity* are almost completely distinct from (not predictive of) home-visitor and observer reports. It is worth re-emphasizing that these item statistics represent the decomposition of *reliable* variation in items. Hence, it is important to consider the reliability of each indicator in conjunction with consistency and method specificity coefficients. For example, although parent-reported positivity explained 27% and 51% of reliable variation in home-visitor- and observer-rated positivity, respectively, there was far more reliable variation to be explained in home-visitor-reported versus observer-rated positivity (R^2 = 0.51 versus 0.11). This helps explain the apparent discrepancy in these values when compared with the bivariate correlations provided in Table 2.

CTC(M–1) models with home visitors as the comparison method—Although the primary focus of this study was on a comparison of parent- with home-visitor-and observer-reported dimensions of positivity and negativity, we also briefly considered key results (model fit and item statistics) when home visitors were considered as the comparison informant/method in the CTC(M–1) model. This analysis also helped us to address the issue of whether the IBR ratings are more similar to parent ratings.[2]

The model with the home visitor as the comparison method was analogous to the third model, resulting in parent- and observer-reported method factors. When first estimated, the residual variance for the parent-reported negativity item was negative, though not significantly different than 0. As such, these residual variance parameters were fixed to 0, and the model was re-estimated. This model fit the data well, as indicated by the likelihood ratio test statistic, $\chi^2(5) = 11.1$, $p = 0.025$, and fit indices (see Table 3).

As is summarized in the second column of Table 4, 86% and 25% of the reliable variation in parent- and observer-rated *positive* behaviour could be explained by home-visitor-reported positive behaviour. The remaining 14% and 75% of reliable variation in parent- and observer-reported positive behaviour was unique to these methods. In contrast, 2% and 39% of the reliable variation in parent- and observer-rated *negative* behaviour could be explained by home-visitor-reported negative behaviour. The remaining 98% and 61% of the reliable variation in parent- and observer-reported negative behaviour was unique to these methods. As noted above, it is important to interpret these statistics with reference to the reliability estimates of each item, which are also reported in Table 4. These results indicate that when only reliable information is considered, home-visitor ratings are moderately predictive of observer ratings

[2]It is important to note that the CTC(M–1) model is not symmetrical (Eid, 2000). This means that model fit will vary as a function of which method is designated as the comparison. As such, to the extent that items are differentially reliable across models, direct comparisons of item statistics (consistency and method specificity coefficients) across models become less meaningful.

for both positivity and negativity. In contrast to parent-rated positivity, home-visitor ratings are not predictive of parent-rated negativity.

CTC(M–1) models with LT coders as the comparison method—For comparative purposes, we analysed a fifth model with observers designated as the comparison method, resulting in parent and home-visitor method factors. When first estimated, the residual variance for the home-visitor rating of positivity item was negative, though not significantly different than 0; therefore, these residual variance parameters were fixed to 0, and the model was re-estimated. This model fit the data well, as indicated by the likelihood ratio test statistic, $\chi^2(4) = 6.7$, $p = 0.15$, and fit indices (see Table 3).

As is summarized in Table 4, 1% and 17% of the reliable variation in parent-and home-visitor-rated *positive* behaviour could be explained by observer-coded positive behaviour. The remaining 99% and 83% of reliable variation in parent-and home-visitor-rated positive behaviour was unique to these methods (informants). Moreover, while 57% and 39% of the reliable variation in parent-and home-visitor-rated *negative* behaviour could be explained by observer-coded negative behaviour, the remaining 43% and 61% of the reliable variation in parent- and home-visitor-rated negative behaviour was unique to these methods. These results indicate that, when only reliable information is considered, observer ratings are not predictive of parent ratings of positivity, are modestly predictive of home-visitor ratings of positivity, and are moderately to strongly predictive of both parent and home-visitor ratings of negativity.

DISCUSSION

The primary goal of the present study was to assess the convergence between parent ratings of infant temperament, the observations of home visitors who observed infant behaviour over an extended period of time, and coders who observed variations in specific behaviours during specific periods of time. Results indicated that whereas there was a moderate degree of convergence between parents and observers with respect to infant positivity, there was little to no convergence across parents and observers with respect to infant negativity.

Previous research interested in how well parents and observers agree on infant temperament has primarily relied on the simple inspection of bivariate correlations (cf. Saudino, Cherny, & Plomin, 2000). Such statistical approaches have shown, fairly consistently, that parents and observers do not agree or show minimal agreement. Even in cases where ratings were aggregated or parents and observers rated the same situations, they continued to show low to modest agreement (Forman *et al*., 2003; Seifer *et al*., 1994, 2004). These results have led to debates about the value of parent ratings and questions about the validity of research that relies solely on parent reports of their own infant's temperament. Indeed, in two separate issues of *Infant Behavior and Development* (Volume 25, 2002, and Volume 26, 2003), temperament researchers argued the pros and cons of using parent-report measures. The methods used by parents and observers to assess temperament, however, may be a significant reason why the findings are so controversial. Comparing parents' perceptions with observers who are trained to focus on specific behaviours recorded during standard laboratory tasks raises questions about the effect of these differing methods on the results. In the present study we attempted to address this concern by using SEM. With this model-fitting procedure we were able to compare and contrast three models, one which assumed that (co)variation would be adequately explained by trait factors (positivity and negativity), one which assumed that (co)variation would be adequately explained by the three different methods (parent ratings, home-visitor ratings, LT coding), and a hybrid model that included both trait and method factors. By allowing systematic variation in both trait and method factors, we were able to examine more precisely whether observers agreed with parent reports of infant temperament. Our results confirmed that the hybrid model was the best-fitting model to the data. Moreover, we found that parents and

observers (regardless of whether the observers used a rating scale applied to 5 h of home observation or coded behaviours from videotape of specific emotion-eliciting tasks) agreed only on the extent to which the infant was positive. When considering only true score variance due to trait and method factors, 51% of the variation in coders' ratings of positivity was explained by parent ratings, whereas 27% of the variation in home-visitor ratings of positivity was explained by parent ratings. Our results suggest some correspondence between parents and observers on infant positive reactivity, but the inclusion of method factors in the model does not appear to improve substantially upon the findings of previous studies when assessing agreement on infant temperament.

Convergence between parents and observers on infant negativity was nonexistent. While this finding is consistent with a number of previous studies, there are several possible explanations for why parents agree with observers on positivity and disagree on negativity. For instance, despite the fact that negative behaviours are highly salient and thus would be easily rated by parents and observers, parents may have downplayed the negativity of their child. This explanation is supported by a study by Seifer *et al*. (2004) in which parents observed their own child and an unfamiliar child during the same tasks as observers and found parents to underrate their own child's, as well as a standard child's, negativity. Secondly, in the present study the home visit and, in particular, the mask and arm restraint tasks were themselves unique and non-typical events, which may have heightened negative affect and artificially inflated any differences between parents and observers.

A third possible explanation is that parents took an idiographic approach (Pelham, 1993) to rating their children. In forming their ratings parents may have taken into consideration their knowledge of their infants' potential for negative reactivity, using a within-person comparison rather than comparing their infants with other infants. For example, a mother may observe her infant to be highly negative to certain situations but non-reactive to other situations. In responding to items that ask about her child's negative reactivity, even when the item specifies a certain situation, the mother may take into account the range of negative reactivity that is possible for her child. Observers, on the other hand, are trained on scales that incorporate the range of all possible degrees of negative reactivity; thus, they use a between-persons comparisons. Even during 5 h of observation, home visitors may not have been exposed to the range of each individual child's negativity.

This reasoning may also explain why parents and observers, particularly coders of the LT tasks, tend to agree on the infants' positive, rather than negative, reactivity. Because of the constraints of the present study, only the approach subscale of the IBQ was available for analysis. Items on the approach scale assess the degree to which the infant gets excited about new toys, places, and persons. Such distinctive characteristics, which would be very salient to parents, may also emerge in tasks that were designed to elicit negative reactivity. In other words, being more positively responsive to the unusual masks more likely maps onto this approach form of positive reactivity than the smiling/laughter subscale of the IBQ. Interestingly, in a study of 12-month-old infants, an excitability/mood factor drawn from infant behaviour during a lab visit was significantly related to positive reactivity as rated by parents (Carnicero *et al*., 2000). This reasoning may also account for the discrepancy between the percentage of variance in home-visitors' and coders' ratings that was explained by parent ratings. Home visitors were explicitly asked to disregard reactivity to the challenge tasks when making their ratings.

The second goal of the present study was to introduce the IBR as an intermediary assessment of infant temperament. We devised the IBR to be more like parent ratings by requiring minimal training and having home visitors base their ratings on a variety of situations including caretaking, mother–infant interaction, and periods in which the infant was not the focus of the parent or the home visitor. To address whether IBR ratings were more like parent ratings, we

re-estimated the SEM CTC(M–1) model using home-visitor ratings as the comparison method. This allowed us to examine the degree to which the IBR ratings explained variance in the parent and observer ratings. The results revealed that the IBR ratings explained the majority (86%) of the reliable (true score) variance in parent ratings of positive reactivity but very little (2%) of the reliable (true score) variance in parent ratings of infant negativity. In comparison, the IBR ratings explained 25% and 39% of reliable (true score) variance in observers' coding of infant positive and negative reactivity. Thus, it appears that if observers are trained minimally on a rating scale and have the opportunity to observe infant behaviours across a number of situations in the home, agreement is improved, but again only for positive reactivity. The majority of the variance in parent ratings of negative reactivity is unique to this method. The restriction of scales used in the IBR for the present study may have contributed to this finding. Further support for the use of the IBR as a parent-like measure is that the IBR showed little convergence with the coded positive behaviour. Interestingly, the results for negative behaviour showed home visitors and observers to agree more than with parents, despite instructing home visitors to ignore the infants' reactions to the challenge tasks when making their ratings. Because the challenge tasks were conducted in the home it may have been difficult for home visitors to ignore. Future researchers may want to compare IBR ratings with observers' coding of behaviour in the laboratory.

In sum, it appears that, when observers are given similar methods to those of parents, as well as opportunities to observe infant behaviour across a number of situations, agreement with parents is relatively good when rating positive behaviours. These findings argue that parents may be more reliable reporters of their infants' temperament than previously believed. However, when rating negative behaviours, parents and observers using the IBR continue to disagree. Future studies assessing infant temperament with the IBR might consider adding scales that map more directly onto parent ratings of infant temperament.

Although the results of the present study confirm earlier studies examining agreement between parents and observers on infant temperament, there are several limitations that suggest caution when interpreting the data. Owing to the constraints of the longitudinal study (e.g. participant burden) only a subset of the scales from the IBQ was used. Thus, a direct measure of parent perceptions of smiling/laughter was not available and the approach subscale was substituted. Likewise, reactions to positive stimuli were not elicited during the home visit but, rather, positive reactivity during the two tasks used to elicit negativity was coded. While convergence was good between the parent ratings and observers' coding of positivity, future research should use parent-rated and observer-rated measures of positive reactivity that assess the same trait. Although currently used as a measure of temperament, the IBR used in the present study was designed for use with a developmental assessment. In addition, the items of the IBR did not assess the same dimensions as the IBQ and the coded behaviours. These differences may have affected the lack of convergence between parent and home visit ratings of infant negativity, in particular. Although our sampling techniques (i.e. over-sampling low-income and African-American families) and analytic strategies increase our confidence in generalizing these data to the broader population, parents of varying income levels or ethnicities may view the temperament of their infants differently thus influencing observer/parent agreement. Plans are currently underway to investigate whether income, as well as race, moderates agreement among methods of temperament assessment.[3]

Whereas the data are encouraging regarding parents as reliable reporters of their infants' temperament, the data also suggest that parent report of their own child's temperament is a relatively idiosyncratic method that may reflect parents' unique perspective. Our use of a SEM model that was developed to facilitate a comparison of informants in the context of multitrait-

[3]We would like to thank one of our reviewers for this suggestion.

multimethod data will not end the debate on whether parent ratings of infant temperament are accurate or inaccurate. However, rather than assuming that parents are poor reporters of their infants' temperament, we concur with Rothbart and Bates (1998) that it would be more realistic to take a components-of-variance approach. Parents, as well as observers, vary in the degree to which subjective and objective factors contribute to their ratings and, as suggested by Bornstein *et al*. (1991), may tap different portions of the variance in infant temperament. In conclusion, we join our many colleagues in suggesting that, when doing research in infant temperament, each method should be considered carefully and, where possible, used in combination. 'Agreeing to disagree', therefore, may be the most conciliatory and productive course of action.

Acknowledgments

This research was supported by a grant from the National Institute for Child Health and Human Development (5PO1-HD-39667), Lynne Vernon-Feagans and Martha Cox, PIs, with co-funding from the National Institute on Drug Abuse. The Family Life Project Key Investigators include Lynne Vernon-Feagans, Martha Cox, Clancy Blair, Peg Burchinal, Linda Burton, Keith Crnic, Nan Crouter, Patricia Garrett-Peters, Doug Granger, Mark Greenberg, Stephanie Lanza, Adele Miccio, Roger Mills-Koonce, Deborah Skinner, Cynthia Stifter, Lorraine Taylor, Emily Werner, and Mike Willoughby.

References

Aksan N, Kochanska G. Links between systems of inhibition from infancy to preschool years. Child Development 2004;75:1477–1490. [PubMed: 15369526]

Arbuckle, JL. Full information estimation in the presence of incomplete data. In: Marcoulides, GA.; Schumacker, RE., editors. Advanced structural equation modeling. Mahwah: Lawrence Erlbaum Associates; 1996. p. 243-277.

Bates J, Bayles K. Objective and subjective components in mothers' perceptions of their children from age 6 months to 3 years. Merrill-Palmer Quarterly 1984;30:111–130.

Bayley, N. Bayley Scales of Infant Development. New York: Psychological Corporation; 1969.

Belsky J, Fish M, Isabella RA. Continuity and discontinuity in infant negative and positive emotionality: Family antecedents and attachment consequences. Developmental Psychology 1991;27:421–431.

Belsky J, Hsieh KH, Crnic K. Infant positive and negative emotionality: One dimension or two? Developmental Psychology 1996;32:289–298.

Bohlin G, Hagekull B, Lindhagen K. Dimensions of infant behavior. Infant Behavior and Development 1981;4:83–96.

Bornstein M, Gaughran J, Sequi I. Multi-method assessment of infant temperament: Mother questionnaire and mother and observer reports evaluated and compared at five months using the infant temperament measure. International Journal of Behavioral Development 1991;14:131–151.

Braungart J, Plomin R, DeFries J, Fulker D. Genetic influence on tester-rated infant temperamental as assessed by Bayle's Infant Behavior Record: Nonadoptive and adoptive siblings and twins. Developmental Psychology 1992;28:40–47.

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.

Diener E, Smith H, Fujita F. The personality structure of affect. Journal of Personality and Social Psychology 1995;69:130–141.

Dill, B. Rediscovering rural America. In: Blau, JR., editor. Blackwell companions to sociology. Malden, MA: Blackwell Publishing; 2001. p. 196-210.

Eid M. A multitrait-multimethod model with minimal assumptions. Psychometrika 2000;65(2):241–261.

Eid, M.; Lischetzke, T.; Nussbeck, FW. Structural equation models for multitrait-multimethod data. In: Eid, M.; Diener, E., editors. Handbook of multilevel measurement in psychology. Washington, DC: American Psychological Association; 2006. p. 283-299.

Eid M, Lischetzke T, Nussbeck FW, Trierweiler LI. Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. Psychological Methods 2003;8(1):38–60. [PubMed: 12741672]

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

Gartstein MA, Rothbart MK. Studying infant temperament via the revised infant behavior questionnaire. Infant Behavior and Development 2003;26:64–86.

Goldsmith H, Campos J. The structure of temperamental fear and pleasure in infants: A psychometric perspective. Child Development 1990;61:1944–1964. [PubMed: 2083507]

Goldsmith H, Gottesman I. Origins of variation in behavioral style: A longitudinal study of temperament in young twins. Child Development 1981;52:91–103. [PubMed: 7195330]

Goldsmith H, Hewitt E. Validity of parental report of temperament: Distinctions and needed research. Infant Behavior and Development 2003;26:108–111.

Goldsmith, H.; Rothbart, M. The laboratory temperament assessment battery: Prelocomotor version 3.0. Madison; WI: 1996.

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.

Kagan, J. Biology and the child. In: Damon, W.; Eisenberg, N., editors. Handbook of child psychology. Vol. 3, social, emotional and personality development. New York: Wiley; 1998. p. 178-235.

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.

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

Mebert C. Dimensions of subjectivity in parent's ratings of infant temperament. Child Development 1991;62:352–361. [PubMed: 2055126]

Meisels S, Cross D, Plunkett J. Use of the Bayley Infant Behavior Record with preterm and full-term infants. Developmental Psychology 1987;23:475–482.

Muthén, LK.; Muthén, BO. Mplus users guide. Vol. 3. Los Angeles, CA: Muthén & Muthén; 2004.

Park S, Belsky J, Putnam S, Crnic K. Infant emotionality, parenting, and 3-year inhibition: Exploring stability and lawful discontinuity in a male sample. Developmental Psychology 1997;33:218–227. [PubMed: 9147831]

Pelham B. The idiographic nature of human personality: Examples of the idiographic self-concept. Journal of Personality and Social Psychology 1993;64:665–677. [PubMed: 8473983]

Putnam S, Gartstein M, Rothbart M. Measurement of fine-grained aspects of toddler temperament: The early childhood behavior questionnaire. Infant Behavior and Development 2006;29:386–401. [PubMed: 17138293]

Putnam S, Stifter C. Behavioral approach-inhibition in toddlers: Prediction from infancy, positive and negative affective components, and relations with behavior problems. Child Development 2005;76:212–226. [PubMed: 15693768]

Rothbart M. Measurement of temperament in infancy. Child Development 1981;52:569–578.

Rothbart M, Ahadi SA, Hershey KL, Fisher P. Investigations of temperament at three to seven years: The children's behavior questionnaire. Child Development 2001;72(5):1394–1408. [PubMed: 11699677]

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

Rothbart M, Evans D, Ahadi S. Temperament and personality: Origins and outcomes. Journal of Personality and Social Psychology 2000;78(1):122–135. [PubMed: 10653510]

Rubin K, Burgess K, Hastings P. Stability and social–behavioral consequences of toddlers' inhibited temperament and parenting behaviors. Child Development 2002;73:483–495. [PubMed: 11949904]

Russell J, Carroll J. On the biopolarity of positive and negative affect. Psychological Bulletin 1999;125:3–30. [PubMed: 9990843]

Sameroff A, Seifer R, Elias P. Sociocultural variability in infant temperament ratings. Child Development 1982;53:164–173. [PubMed: 7060419]

Satorra, A.; Bentler, PM. A scaled difference chisquare test statistic for moment structure analysis. Los Angeles: UCLA Statistics Series; 1999. (paper # 260)

Saudino K, Cherny S, Plomin R. Parent ratings of temperament in twins: Explaining the “too low” dz correlations. Twin Research 2000;3:224–233. [PubMed: 11463143]

Saudino K, Plomin R, DeFries J. Tester-rated temperament at 14, 20 and 24 months: Environmental change and genetic continuity. British Journal of Developmental Psychology 1996;14:129–144.

Seifer R. What do we learn from parent reports of their children’s behavior? Commentary on Vaughn et al.’s critique of early temperament assessments. Infant Behavior and Development 2002;25:117–120.

Seifer R, Sameroff A, Barrett L, Krafchuk E. Infant temperament measured by multiple observations and mother report. Child Development 1994;65:1478–1490. [PubMed: 7982363]

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.

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

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. Developmental 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.

Watson D. Intra-individual and interindividual analyses of positive and negative affect. Journal of Personality and Social Psychology 1988;54:1020–1030. [PubMed: 3397861]

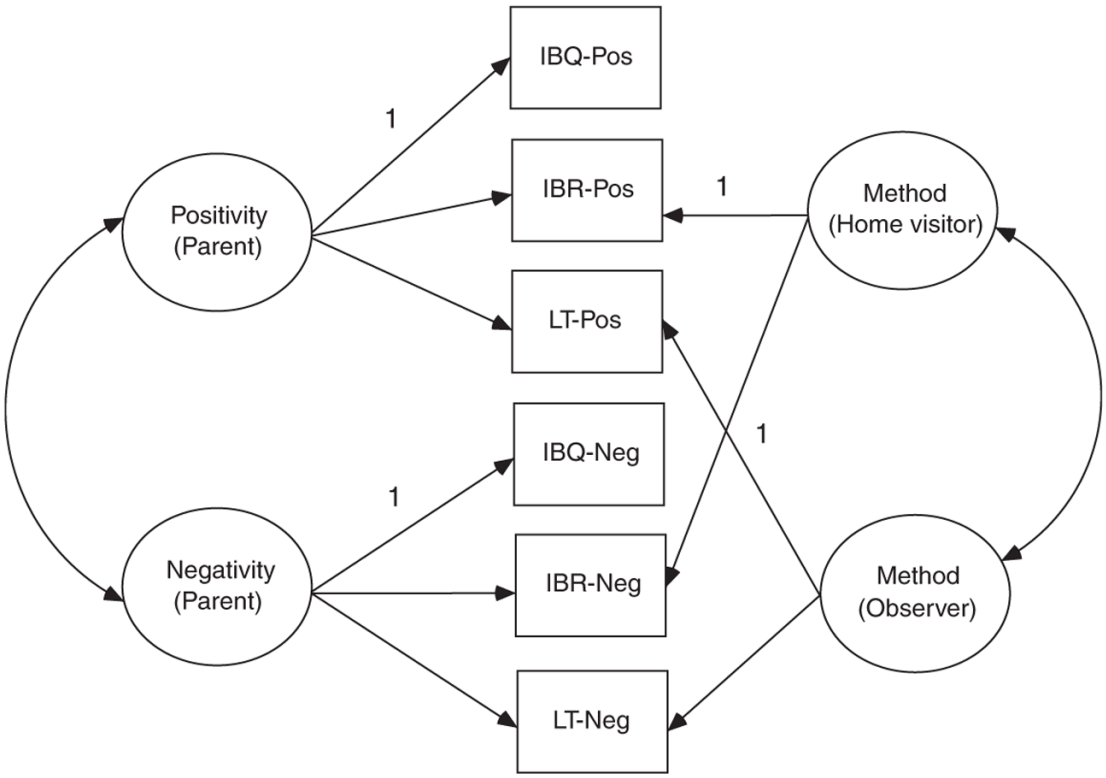

Figure 1.
Correlated trait correlated method–1 model with parent ratings as reference method.

Table 1

Sample description

Variable	Categories/unit	%
State	NC	55
Primary caregiver race	Caucasian	63
Primary caregiver education	Less than high school	19
	GED/high school+	65
	4-Year college degree+	16
Target child race	Caucasian	61
Target child sex	Male	52
		M (S.D.)
Primary caregiver age	Years	26.5 (6.1)
Target child age	Months	7.3 (0.9)
Household income	Income/needs ratio	1.9 (1.8)

Note: M, mean; S.D., standard deviation; 100% of primary caregivers were female and 99.6% of primary caregivers were the biological mother of the target child. The remaining four caregivers consisted of one foster parent, two grandparents, and one other adult relative.

Table 2

Bivariate correlations

Method	Item	1	2	3	4	5	6
Parent	1. IBQ-pos.	1.000					
	2. IBQ-neg.	− 0.033	1.00				
Home visitor	3. IBR-pos.	0.101	0.127	1.000			
	4. IBR-neg.	− 0.059	0.102	−0.603	1.000		
Observer	5. LT-pos.	0.022	−0.105	0.122	−0.069	1.000	
	6. LT-neg.	0.051	0.093	−0.118	0.249	−0.11	1.000

Note: Unweighted coefficients were computed using pairwise deletion methods. Underlined correlations are significant at the <0.001 level. Ns from 837 to 953; IBQ, Infant Behavior Questionnaire; IBR, Infant Behavior Record; LT, LAB-TAB.

Table 3

Synopsis of positive–negative model fit

Model	Description	Comparison informant	χ (df)	CFI	RMSEA	SRMR
1	CT	NA	34.8 (10)	0.95	0.05	0.039
2	CM (only)	NA	34.6 (6)	0.95	0.07	0.028
3	CTC(M–1)	Parent[a]	11.5 (5)	0.99	0.04	0.022
4	CTC(M–1)	Home visitor	11.1 (4)	0.99	0.04	0.019
5	CTC(M–1)	Observer	6.7 (4)	1.0	0.03	0.016

Note: N = 955; CT, correlated trait; CM, correlated method; CTC(M–1), correlated trait correlated method minus one; NA, not applicable; CFI, comparative fit index; RMSEA, root mean-squared error of approximation; SRMR, standardized root mean residual.

[a]Depicted in Figure 1.

Table 4

Item statistics

Construct	Informant	Item	Comparison informant/method								
			Parent (IBQ)			Home visitor (IBR)			Observer (LAB-TAB)		
			R^2	CC	MS	R^2	CC	MS	R^2	CC	MS
Positive	Home visitor	IBR-pos.	0.51	0.27	0.73	0.85	1.00	0.00	1.00	0.17	0.83
Positive	Observer	LT-pos.	0.11	0.51	0.49	0.11	0.25	0.75	0.11	1.00	0.00
Positive	Parent	IBQ-pos.	0.05	1.00	0.00	0.02	0.86	0.14	0.20	0.01	0.99
Negative	Home visitor	IBR-neg.	1.00	0.01	0.99	0.86	1.00	0.00	0.49	0.39	0.61
Negative	Observer	LT-neg.	0.28	0.05	0.95	0.18	0.39	0.61	0.32	1.0	0.00
Negative	Parent	IBQ-neg.	1.00	1.00	0.00	1.00	0.02	0.98	0.08	0.57	0.43

Note: CC, consistency coefficient = amount of true variance for an observed variable that is explained by respective trait factor; MS, method specificity = amount of true variance for an observed variable that is explained by respective method factor; R^2, reliability; IBQ, Infant Behavior Questionnaire; IBR, Infant Behavior Record; LT, LAB-TAB.