Development and Assessment of Short and Very Short Forms of the Infant Behavior Questionnaire–Revised

By: Samuel P. Putnam, Amy L. Helbig, Maria A. Gartstein, Mary K. Rothbart, Esther Leerkes

Putnam, S.P., Helbig, A.L., Gartstein, M.A., Rothbart, M.K., Leerkes, E. (2014). Development and Assessment of Short and Very Short Forms of the Infant Behavior Questionnaire–Revised. *Journal of Personality Assessment, 96(4),* 445-458. doi: 10.1080/00223891.2013.841171

Abstract:

Using data from parents of 761 infants from 6 independent samples, short (91 items, 14 scales) and very short (37 items, 3 broad scales) forms of the Infant Behavior Questionnaire–Revised (IBQ–R), a well-established caregiver report measure of temperament for infants aged 3 to 12 months, were developed. The forms were subsequently evaluated with data from 1,619 participants from 11 samples. Over 90% of Cronbach's alphas and part–whole correlations calculated for the short and very short form scales were greater than.70. Interparent agreement was nearly identical to that obtained with standard IBQ–R scales, averaging.41 and ranging from.06 to.76. Longitudinal stability over multiple time spans, and estimated retest reliability of the short form scales, were highly similar to those of standard forms, with estimated retest reliability averaging.72 and ranging from.54 to.93. Convergent and predictive validity of select short form scales were comparable to, but slightly lower, than those observed for standard IBQ–R scales. Recommendations for the use of the standard, short, and very short scales are discussed.

Keywords: Infant Behavior | Child Development | Interparent Agreement | Infant Temperament

Article:

Research on infant temperament has evolved substantially in recent years. Classic investigations concerned basic issues such as establishment of standardized measures (e.g., Carey & McDevitt, 1978; Rothbart, 1981), demonstration of longitudinal stability (e.g., Rothbart, 1986), quantifying levels of heritability (e.g., Buss & Plomin,1984), identifying physiological correlates (e.g., Kagan, Reznick, & Snidman, 1987), and determining socially meaningful outcomes (e.g., Thomas & Chess, 1977) associated with early behavioral predispositions. The success of these endeavors led to wide acceptance of the importance of temperament (Zentner & Shiner, 2012). Subsequently, temperament has been increasingly incorporated into research in which individual differences are not a primary focus, but are instead assessed alongside a wide range of variables expected to interact as predictors of adaptive and maladaptive outcomes (see Rothbart, 2011).

Early temperament research also frequently considered a restricted set of dimensions, such as emotionality, activity, and sociability (e.g., Buss & Plomin, 1984), or assessed constructs with considerable conceptual overlap, using scales with limited internal consistency (e.g., Carey & McDevitt, 1978; see Rothbart & Mauro, 1990). In contrast, contemporary approaches have expanded the list of meaningful temperament dimensions, incorporating constructs gleaned from neuroscience, adult temperament research, and investigations of nonhuman animals (e.g., Gartstein & Rothbart, 2003). More recent examples of temperament research, grounded in the psychobiological framework (Rothbart & Bates, 2006), have also emphasized aspects of regulation and recommended decomposition of broad traits, such as emotionality, into more nuanced elements, such as susceptibility to different negative emotions and differing levels of stimulation required to elicit positive affect (e.g., Rothbart, Ahadi, Hershey, & Fisher, 2001). Accumulating evidence has also demonstrated meaningful differences in the associations of these fine-grained traits with a variety of important childhood outcomes (e.g., behavior problems; Gartstein, Putnam, & Rothbart, 2012).

These two directions in the evolution of the field have led to an interesting conflict. As the temperament domain includes a greater number of characteristics, there are increases in the time and effort required of research participants for their assessment. When temperament represents only a portion of the data collection, researchers are also limited by the demands they are able to place on participants in their assessment of temperament. The goal of the current effort is to resolve this conflict through the creation of two new versions of the Infant Behavior Questionnaire–Revised (IBQ–R; Gartstein & Rothbart, 2003), a parent-report instrument that assesses 14 fine-grained aspects of temperament in infants between 3 and 12 months of age. In addition to a short form, we constructed a very short form, which can be used to measure three broad dimensions that have emerged from explorations of the structure of the IBQ–R. The primary goal of the analyses described herein is to investigate the psychometric characteristics of the abbreviated scales to provide comparison to these characteristics in the original scales.

The IBQ–R, and its predecessor, the Infant Behavior Questionnaire (IBQ; Rothbart, 1981), are based in a definition of temperament as constitutionally based individual differences in reactivity and self-regulation, influenced over time by heredity and experience (Rothbart & Bates, 2006). The original IBQ contained scales to assess six aspects of temperament: Activity Level, Fear, Distress to Limitations, Smiling and Laughter, Soothability, and Duration of Orienting. The revised instrument was created in response to developments in understanding of temperament and contained eight new scales (Approach, Vocal Reactivity, High and Low Intensity Pleasure, Perceptual Sensitivity, Sadness, Falling Reactivity, and Cuddliness), as well as minor revisions of the original six scales (Gartstein & Rothbart, 2003), resulting in a 191-item measure of 14 scales of 10 to 17 items each. Both instruments contain items that were rationally generated based on conceptual definitions for each scale.

To minimize parental biases associated with poor recall, making abstract evaluations, or comparative judgments, parents are asked to report, on a 7-point scale, the frequency with which infants have enacted specific behaviors in common situations during the past week or 2 weeks. Parents are also provided with a “not applicable” response option for use when the child has not been observed in the situation described. Since their introductions, the IBQ and IBQ–R have been among the most frequently used measures of infant temperament. Supporting their validity, the article concerning the development of the IBQ (i.e., Rothbart, 1981) has been cited more than 383 times, and the article regarding the creation of the IBQ–R (i.e., Gartstein & Rothbart, 2003) has been cited 142 times (PsycInfo, March 21, 2013), including studies demonstrating convergent validity of these instruments with observational measures (Gartstein et al., 2010; Gartstein & Marmion, 2008; Kochanska, Coy, Tjebkes, & Husarek, 1998; Parade & Leerkes, 2008).

We wished to create abbreviated scales that approximated the full content of the original scales, yet still demonstrated acceptable internal consistency. These two goals are often difficult to reconcile because items most closely correlated with one another (contributing to internal consistency) might be redundant in content, narrowing the breadth of the measured construct, a phenomenon known as the attenuation paradox (Loevinger, 1954). We balanced these two goals by basing our decisions regarding item retention not only on the degree to which items contributed to internal consistency, but also referred to factor analyses of the individual scales to ensure that all facets of a given trait were represented, and closely examined the content of items to minimize repetition in the substance of the questions.

We also wished to maximize the generalizability of the scales to multiple samples, including those assessing temperament at different ages, with mothers and fathers serving as informants. A common “sin” of short form development concerns basing item inclusion decisions on a single sample, a practice that leads to overestimates of the degree of internal consistency relative to indexes obtained in other samples (Smith, McCarthy, & Anderson, 2000). This concern is particularly salient for developmental research. Due to the rapid pace of development during infancy, behaviors associated with an element of temperament early in the first year of life might not be strong markers of the trait at older ages. To address these issues, analyses considered in the creation of the short and very short forms were carried out on data gathered from six separate samples of children, including two samples for which both mother and father reports were obtained, covering the age range of 3 to 12 months, for which the IBQ–R was designed.

The very short form was not developed to capture the 14 fine-grained scales contained in the original IBQ–R, but rather to measure factors that have been derived from exploratory factor analyses of the instrument. Three broad components of the IBQ–R—Negative Emotionality (NEG), Positive Affectivity/Surgency (PAS), and Orienting/Regulatory Capacity (ORC)—have emerged across studies in the United States and elsewhere (e.g., Gartstein, Knyazev, & Slobodskaya, 2005), and bear strong similarity to those obtained with fine-grained temperament measures in older children and adults (Evans & Rothbart, 2007; Putnam, Ellis, & Rothbart, 2001). The Positive Affectivity/Surgency (PAS) factor is made up of Approach, Vocal Reactivity, High Intensity Pleasure, Smiling and Laughter, Activity Level, and Perceptual Sensitivity, with all six scales demonstrating strong primary loadings on this factor, roughly similar to the personality dimension of Extraversion. A second factor, Negative Affectivity (NEG), is analogous to the personality trait of Neuroticism, and is characterized by high positive loadings on Sadness, Distress to Limitations, and Fear, as well as high negative loadings on Falling Reactivity. A final factor, labeled Orienting/Regulatory Capacity (ORC), is defined by Duration of Orienting, Low Intensity Pleasure, Cuddliness, and Soothability; it has been shown to predict later emerging Effortful Control (Putnam, Rothbart, & Gartstein, 2008), in turn linked with the adult personality trait of Conscientiousness (Rothbart, Ahadi, & Evans, 2000).

Following the construction of the short and very short forms, several steps were taken to assess the psychometric properties of these instruments. In addition to calculating the internal consistency of scores from the short form scales and corrected standard-short form correlations, we assessed correspondence between maternal and paternal ratings and longitudinal rank-order stability. In consideration of arguments made emphatically by Thompson (e.g., 1994; Vacha-Haase & Thompson, 2011) and others (e.g., Streiner, 2003; Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999) that reliability and validity are not properties of a test, per se, but of scores that might fluctuate substantially when a given instrument is used with different samples, our analyses have been conducted across multiple data sets independently, rather than aggregating across samples, to more accurately estimate the psychometric properties of scores generated in future administrations of the measures.

In summary, the purpose of these studies was to develop and assess short and very short forms of parent-report measures of temperament for children between 3 and 12 months of age. Statistical and theoretical considerations were taken into account to make item-inclusion decisions, as were issues of comparability across age. We first describe the samples and procedures used to make decisions regarding item retention in Study 1, which also contains assessment of interrater agreement. Subsequently, internal consistency, corrected part–whole correlations, longitudinal rank-order stability, estimated retest reliability, and cross-informant correlations are presented in Study 2, calculated from a number of samples that had completed the standard form. To diminish the possibility that psychometric qualities of short form scores extracted from data collected with the long form would not be retained when the abbreviated measures were administered, Study 3 involves analyses of internal consistency of data from two samples in which participants completed the short form, one sample who completed the very short form, and a sample who completed a hybrid version consisting of the very short form and items from select short form scales. Finally, Study 4 contains comparisons of previously reported findings concerning convergent and predictive validity of standard IBQ–R scales with those obtained when these findings were replicated with short form scales.

Study 1: Scale Construction and Assessment of Cross-Rater Agreement Samples

Electronic mail correspondence was sent to all individuals from English-speaking countries who had requested information on, or access to, the IBQ–R between 2004 and 2008. Responses to this e-mail yielded data sets collected by four principal investigators from eight different reporters on six separate cohorts of children (overall child *n* = 761; 380 female). Two samples were rated only by primary caregivers at a single time point. The first of these was collected by Ken Ong at Addenbrooke's Hospital, Cambridge, UK, and included mothers' ratings of 154 infants (73 females) at an average age of 3.49 months (*SD* = 2.99, range = 2.56–7.91 months). As indicated in de Lauzon-Guillain et al. (2012), the average maternal age was 33.2 years (*SD* = 4.8). Of parents reporting their education, 8% had completed O-level education (typically completed at approximately age 16, at the end of compulsory education), 14% had completed A-levels (typically completed after completion of preuniversity schooling around age 18), and 81% completed degree-level studies. The next data set was collected by Maria Gartstein of Washington State University and included mother reports on 146 children (73 females) at an average age of 7.59 months (*SD* = 14.49, range = 2.56–12.56 months). As described in Gartstein et al. (2010), the mothers in this sample were primarily White (92%), ranged in age from 20 to 46 (*M* = 29), and were highly educated, with 97% completing high school and 65% completing a bachelor's degree. Gartstein also contributed a second data set for which children were rated at 4, 6, 8, 10, and 12 months of age by both mothers, *n* = 135 (67 females), and fathers, *n* = 72 (32 females). As described in Gartstein et al. (2010), at the initial collection, mothers were primarily White (92%), their ages ranged from 20 to 46 (*M* = 30); 99% had completed high school, with 32% having also completed a bachelor's degree. Fathers were also primarily White (92%); their ages ranged from 20 to 45 (*M* = 30); 97% had completed high school, and 35% had completed a bachelor's degree. Martha Ann Bell of Virginia Tech University contributed primary caregiver-report data regarding two cohorts of children, each rated by mothers at both 5 and 10 months child age. As described by Morasch and Bell (2012), the first sample, *n* = 106 (56 female), was primarily White, with two African American, one Asian, one "other," and nine multiracial infants. All parents in this sample had completed high school, and 71% had a college degree. Mothers' ages ranged from 20 to 38 years (*M* = 30) and fathers' ages ranged from 23 to 52 (*M* = 33). The racial composition of the second sample (*n* = 103; 59 female), was 92% White, 8% multiracial, and 1% Asian. Ninety-eight percent of mothers and 97% of fathers in this sample completed high school, with 70% of mothers and 62% of fathers also completing college. Esther Leerkes of the University of North Carolina at Greensboro collected data on a single cohort of children at 6 months from both mothers, *n* = 117 (52 female infants), and fathers, *n* = 79 (33 female infants). As described by Parade and Leerkes (2008), mothers' ages ranged from 15 to 38 (*M* = 28), and fathers' ages ranged from 21 to 43 (*M* = 31); 67% of both mothers and fathers had college degrees; 77% of mothers and 84% of fathers were White; and family income ranged from $6,000 to $190,000 (*M* = $70,000).

We wished to utilize the richness of the longitudinal data sets (i.e., those from Gartstein and Bell), but did not want them to weigh more strongly in our decisions than data from those cohorts measured at a single time point. To this end, for each cohort the item–total correlations described next were calculated at each age, with the resulting coefficients averaged. Conversely, because we wanted to ensure that items selected worked well for mothers and fathers, analyses of internal consistency were conducted separately for mothers and fathers. Analyses of internal consistency, therefore, were conducted on data sets representing eight reporters across the six cohorts (Ong, Gartstein 1, Gartstein 2 mothers, Gartstein 2 fathers, Bell 1, Bell 2, Leerkes mothers, Leerkes fathers).

For the scale-level exploratory factor analyses constituting our second phase of instrument construction, we wished to maximize the ratio of subjects to variables, so data from all cohorts were combined. For the two cohorts with data from both parents, mother reports were used because they contained a lower proportion of missing cases. For cohorts measured more than one time (i.e., Bell data set 1, Bell data set 2, and Gartstein data set 2), cases were selected with the goal of a combined data set relatively equally distributed across ages 3 to 12 months. Specifically, only 5-month data were used from the Bell 2 data set; only 10-month data were used from the Bell 1 data set; and the Gartstein 2 data set was randomly split into subsets of 67 and 68 cases for which the 8- and 12-month data, respectively, were used. Due to a number of missing cases at these time points in the Gartstein 2 data set, the number of included participants was lower, *n* = 46 at 8 months and *n* = 50 at 12 months.

Procedure

Short form

Scale construction procedures were modeled on those used by Putnam and Rothbart (2006) in the creation of the Children's Behavior Questionnaire (CBQ) short and very short forms. Following the Putnam and Rothbart procedure, we first calculated item–total correlations for each scale, separately for each of the eight data sets, subsequently averaging these item–total correlations over the eight groups. The six items with the highest mean item–total correlations were then used to form working scales.

We desired a minimum alpha of.65 for every scale in each data set. Although.70 is widely considered a cutoff point for acceptable internal consistency (George & Mallery, 2003; Nunnally, 1978), this cutoff point has also been criticized as arbitrary (e.g., Goodwin & Goodwin, 1999; Knapp & Brown, 1995). Because several scales were multidimensional, as described later, we were concerned that requiring this standard across multiple data sets might unnecessarily limit the conceptual breadth of the short scales. For eight of the working scales, alpha exceeded.65 in each data set. Four scales (Activity Level, Distress to Limitations, Smiling and Laughter, and High Intensity Pleasure) generated alphas >.65 in one group and two (Vocal Reactivity and Soothability) were below the threshold in two data sets. By adding a single item to each of these six working scales, we were able to raise internal consistency to acceptable levels in all data sets for all scales except Smiling and Laughter; that is, adding items to the Smiling and Laughter scale did not raise alpha appreciably in the data set for which it performed poorly (Gartstein 1). Because alpha was over.77 for the original six-item working scale in all other data sets, we retained this six-item scale for the short form. Thus, the short form contains nine six-item scales and five seven-item scales.

To combat attenuation of the scales resulting in excessively narrow content, we then conducted item-level principal axis factoring on each scale of the original IBQ–R. The appropriate number of factors to extract and rotate was determined through the mathematical analog to visual scree analysis developed by Zoski and Jurs (1996). This technique yields results that are largely consistent with others commonly used for this purpose, such as parallel analysis or minimum average parcel methods, but tends to identify more factors than other procedures (Canivez & Watkins, 2010a, 2010b). Because our goal was to reflect the content of the longer scales, we felt that oversampling of factors might be beneficial for identifying important item content to retain for the shortened scales. When more than one factor appeared nontrivial, factors were subjected to the oblimin rotation to identify the items associated with each factor, comparing the output from these analyses to the working scale items. When all factors were not represented equally in the working scale, items were replaced with those reflecting the newly identified factor structure most adequately, deemed on the basis of factor loading magnitudes. Alpha coefficients were then calculated on these revised working scales. When the revised working scales did not yield alphas >.65 for all data sets, alternative items contributing strongly to internal consistency in the troublesome data sets were tested in an iterative fashion. For example, four factors were apparent in the Fear scale: the first including items about meeting unfamiliar adults, the second about being approached by others, the third about sudden changes in the environment, and the fourth concerning new people entering the home. The initial working scale included all items from the first factor, and three of these were replaced with items from the other factors. When alpha was calculated on this new scale, the alpha for one data set fell below.65. In this data set, one item from Factor 1 demonstrated low item-interitem correlations, and was replaced by an item from Factor 4. The resulting scale demonstrated alphas above.65 in all data sets.

Six of the scales (Distress to Limitations, Duration of Orienting, Smiling and Laughter, High Intensity Pleasure, Low Intensity Pleasure, and Vocal Reactivity) were revealed to be unidimensional, whereas the others were characterized by two to five factors. For several of these (Fear, Falling Reactivity, Approach, Cuddliness, Perceptual Sensitivity), all factors were represented in the initial working scales, or it was possible to remove and replace items to reflect all facets of the intended dimension while maintaining adequate levels of internal consistency in all samples. However, it was not possible to represent all factors for three of the IBQ–R scales. Three factors emerged from Activity Level, reflecting behavior during daily care, during sleep, and struggling in response to restraint. It was not possible to include items regarding activity during sleep while keeping alphas over.65 in all data sets, and this facet was not included on the short scale. The Sadness scale contained five factors, and one of these, with items regarding infants' empathetic responses to sad others, was not retained for the short scale. Four factors characterized the Soothability scale. Two of these factors, one including items indicating the likelihood that the infant would not soothe immediately, but would within the first 2 minutes; and another regarding infants' soothing tendencies when given a toy, were not retained.

In a final phase, all scales were inspected with respect to breadth of item content. When more than one item in a working scale referred to very similar child behaviors or very similar contexts, we attempted to replace it with an item judged by the first two authors not to overlap in content with other working scale items. For example, the Perceptual Sensitivity working scale contained three items concerning the infant's reactions to soft sounds and only one item referring to reactions to texture, and it was decided to omit a sound item, replacing it with a texture question. As with decisions made on the basis of the factor analyses, this step was carried out iteratively with the goal of maintaining alphas >.65 in all data sets. Further details regarding the item selection process for all scales are available on request from the corresponding author.

Very short form

Construction of the very short form was carried out following the completion of the short form. In addition to choosing items that correlated highly with their intended factor, we sought scales that were relatively orthogonal, and thus selected items that did not correlate with the other two factors. To arrive at an index representing this dual intent for each item, we first calculated factor scores for PAS, NEG, and ORC by averaging standard scale scores corresponding to the factor. Next, the correlations between each short form item and these three factors were calculated for each data set, and the absolute value of the correlation coefficients (averaged across the eight data sets) for the two "nontarget" factors and the item were averaged and subtracted from the coefficient between the item and the target factor (averaged across the eight data sets). Items with high values for this index were then considered for 12-item working scales consisting of equal numbers of items from each fine-grained scale associated with the factor, with the content of the individual items taken into consideration to avoid overlap. Alpha was then calculated for each working scale on each data set. As with the construction of the short form, if alpha was <.65 for any data set, items detracting from internal consistency in that data set were replaced in an iterative fashion until alphas >.65 were achieved for all data sets. For the PAS scale, an additional item from the Activity Level scale was included to bring reliability above.65 for all sets, such that the final instrument contains two 12-item scales and one 13-item scale. In addition, although Falling Reactivity typically loads on NEG, no Falling Reactivity short form items were sufficiently highly correlated with the Negative Affect factor score to warrant their inclusion in the very short form, and this very short form scale only contains items from Sadness, Distress to Limitations, and Fear.

Table 1 Interparent agreement of Infant Behavior Questionnaire–Revised standard and short form scales, and very short form factor scores, in Study 1.

	Gartstein Sample						**Leerkes**
							Sample
Scale	**4 Months**[a]	**6 Months**[b]	**8 Months**[c]	**10 Months**[d]	**12 Months**[e]	**Average**	**6 Months**[f]
Activity Level							
Standard	.45	.54	.54	.48	.50	.50	.43
Short	.46	.49	.55	.42	.42	.47	.46
Approach							
Standard	.36	.40	.36	.51	.38	.40	.48
Short	.40	.39	.34	.46	.40	.40	.47
Cuddliness							
Standard	.28	.40	.49	.23	.34	.35	.14
Short	.38	.39	.43	.28	.34	.36	.12
Distress to Limitations							
Standard	.33	.38	.62	.59	.68	.52	.30
Short	.32	.38	.63	.44	.56	.47	.43
Duration of Orienting							
Standard	.39	.43	.26	.50	.45	.41	.32
Short	.38	.29	.36	.43	.40	.37	.26
Falling Reactivity							
Standard	.49	.52	.60	.51	.38	.50	.40
Short	.56	.46	.59	.52	.53	.53	.31
Fear							
Standard	.40	.70	.70	.58	.72	.62	.28
Short	.33	.65	.76	.46	.71	.58	.26

High Intensity Pleasure							
Standard	.49	.55	.34	.50	.37	.45	.11
Short	.51	.54	.26	.58	.38	.45	.21
Low Intensity Pleasure							
Standard	.43	.38	.36	.67	.58	.48	.30
Short	.36	.39	.31	.60	.58	.45	.29
Perceptual Sensitivity							
Standard	.28	.30	.45	.19	.19	.28	.30
Short	.26	.27	.39	.23	.23	.28	.31
Sadness							
Standard	.34	.36	.28	.27	.28	.31	.27
Short	.27	.36	.28	.20	.24	.27	.29
Smiling and Laughter							
Standard	.48	.50	.41	.58	.54	.50	.43
Short	.50	.44	.37	.58	.49	.48	.43
Soothability							
Standard	.29	.33	.30	.27	.18	.27	.08
Short	.24	.49	.28	.44	.15	.32	.06
Vocal Reactivity							
Standard	.49	.44	.37	.46	.48	.45	.27
Short	.54	.48	.42	.43	.57	.49	.25
VSF PAS	.61	.49	.49	.52	.56	.53	.45
VSF NEG	.36	.49	.55	.35	.54	.46	.36
VSF ORC	.32	.37	.39	.43	.43	.39	.28

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 68. [b]*n* = 64. [c]*n* = 55. [d]*n* = 43. [e]*n* = 49. [f]*n* = 79.

Results (Interrater Agreement)

Due to space considerations, and because internal consistency is expected to be favorably biased in the construction samples (Kopalle & Lehmann, 1997), we do not report alphas, standard-to-short-form correlations, or longitudinal stability coefficients for the Study 1 samples. Because we were unable to obtain data with multiple informants for Studies 2 and 3, we report cross-informant agreement for the relevant Study 1 data sets only. For these analyses, and all reported for Studies 2 and 3, item-level missing data were replaced on the basis of maximum likelihood estimation, using the expectation maximization algorithm, as Enders (2004) suggested this method yields more accurate estimates of reliability than other practices. Although the interparent agreement of the standard form scales in the Leerkes data was previously published by Parade and Leerkes (2008), we present them here as well to facilitate comparisons with the correlations obtained for the short form scales.

As shown in Table 1, the degree of parental agreement was highly similar for the standard and short form scales at all ages and across both data sets. Consistent with results obtained by Gartstein and Rothbart (2003) using the standard IBQ–R, parents agreed particularly strongly regarding their infants’ Fear, Distress to Limitations, and Activity Level, whereas relatively low agreement was obtained for Soothability and Perceptual Sensitivity, especially during older ages in the Gartstein data; and for High Intensity Pleasure, Soothability, and Cuddliness in Leerkes's data. Regarding the very short form, parental agreement was consistently highest for PAS, and less robust for ORC.

Discussion

Because interrater agreement was not a criteria used for selection of items on the short and very short forms, the fact that parental agreement for the abbreviated measures approximated, and for some scales surpassed, levels demonstrated with the standard form inspires confidence in the shortened instrument. Due to the rapid pace of development between 4 and 12 months of age, it is additionally reassuring that the degree of interparent agreement was largely consistent across multiple age points. The levels of consistency between parents for most scales are similar to those demonstrated for other temperament measures (see review by Slabach, Morrow, & Wachs, 1991), including standard and abbreviated scales of the CBQ (Putnam & Rothbart, 2006; Rothbart et al., 2001), which served as a model for the development of the infant measures assessed in this study.

Differential agreement of individual scales is also consistent with the results of analyses of the CBQ, in that parents often did not agree with respect to their ratings of Perceptual Sensitivity, suggesting a degree of subjectivity in parents’ ratings due to the subtlety of behaviors indicating perceptual awareness, in comparison to more readily observed indexes of other traits. The Soothability and Cuddliness scales, in both their standard and short forms, also exhibited interparent agreement that was substantially lower than that found for other scales. Gartstein and

Rothbart (2003) suggested that low interparent agreement for the Soothability scale reflected differences in the effectiveness of soothing behaviors when enacted by different parents. Similarly, low interobserver agreement of the Cuddliness scale might indicate differences between parents in the emerging relationships each has formed with their infant (Parade & Leerkes, 2008). Because the ORC scale of the very short form is partially comprised of items from the Cuddliness and Soothability scales, it is not surprising that this scale demonstrated lower interparent agreement, relative to the PAS and NEG scales.

Study 2: Short and Very Short Forms Extracted From Standard Form Data Samples

Data sets were acquired by e-mailing researchers who had requested the IBQ–R or published research using the instrument between 2006 and 2011, and were obtained from the following sources:

Susan Calkins at University of North Carolina, Greensboro: 195 (92 female; 114 White, 49 African American, 17 Hispanic, and 15 multiracial) infants assessed at 5 months of age, with 191 of these providing IBQ–R data. Ninety-six percent of mothers and 93% of the fathers who reported educational information had at least a high school diploma. Forty-six percent of the mothers had a college degree, and 18% had an advanced graduate degree. Thirty-three percent of fathers had a college degree, and 16% had an advanced degree. Mothers were approximately 29.1 years old at the time of the child's birth (range = 14–42) and fathers were approximately 31.8 years old (range = 18–58).

Julia Braungart-Rieker at Notre Dame University: A longitudinal sample of infants assessed at 3 (*n* = 131), 5 (*n* = 127), 7 (*n* = 116), 12 (*n* = 116), and 14 (*n* = 106) months. As described by Braungart-Rieker, Hill-Soderlund, and Karrass (2010), the original sample of 143 mothers recruited for this study were primarily White (94%), ranged in age from 17 to 43 years (*M* = 29), and reported annual family incomes ranging from $10,000 to $150,000 (median = $45, 000). The large majority (95%) had completed high school, and 42% had also completed college.

Stephen Porges at University of Illinois, Chicago: 119 infants (68 female) ranging in age from 2.55 to 12.5 months (*M* = 7.00, *SD* = 2.59). Mothers’ ages ranged from 19 to 42 (*M* = 30), all but five had completed high school, and 61 had also completed college. Forty-nine percent of the infants were White, 40% were African American, 2% were Asian, and 8% were multiracial.

Maria Gartstein at Washington State University: 68 infants (32 female) ranging in age from 25 to 59 weeks (*M* = 40.18, *SD* = 10.53). As described by Gartstein and Marmion

(2008), this sample was primarily White (82%) and Asian American (9%), with 4% Latino, 3% Filipino, and 2% African American families.

Shannon Ross-Sheehy at University of Iowa: 54 infants (29 female) assessed at 4 months of age.

Elysia Poggi Davis at University of California–Irvine: 223 infants (105 female) assessed at 3 months of age.

Demographic information other than child age was not available for the latter two samples.

Results

Descriptive statistics

Scale scores for the long, short, and very short forms were calculated as the mean of scale items (after reverse-scoring items when necessary, for instance, “When rocked or hugged, how often did your baby seem eager to get away” on the Cuddliness scale). Means and standard deviations for all samples are found in Table 2.

Table 2 Means and standard deviations of Infant Behavior Questionnaire–Revised standard, short, and very short form scales in Study 2.

	Braungart-Rieker[a]		**Calkins**[b]		**Davis**[c]		**Gartstein**[d]		**Ross-Sheehy**[e]		**Porges**[f]	
Scale	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Activity Level												
Standard	4.32	.76	4.52	.76	3.63	.79	4.44	.93	3.91	.71	4.32	.98
Short	3.96	.92	4.19	.94	3.29	.94	4.07	1.05	3.51	.86	4.00	1.19
Approach												
Standard	4.80	.81	5.06	.91	3.66	1.13	5.48	.79	3.92	.98	5.24	1.09
Short	4.71	.90	5.00	1.08	3.47	1.22	5.46	.88	3.66	1.15	5.15	1.26
Cuddliness												
Standard	5.48	.65	5.73	.66	6.02	.60	5.53	.74	5.96	.46	5.71	.69

Short	5.55	.74	5.86	.80	6.23	.69	5.50	.89	6.19	.46	5.79	.85
Distress to Limitations												
Standard	3.72	.74	3.59	.86	3.39	.77	3.81	.96	3.33	.80	3.63	.90
Short	3.96	.98	3.89	1.06	3.66	1.02	3.87	1.14	3.57	.99	3.83	1.12
Duration of Orienting												
Standard	3.85	.97	4.26	1.03	3.84	1.12	3.62	.96	3.69	.99	4.01	1.04
Short	3.67	1.09	4.04	1.11	3.62	1.21	3.33	1.06	3.31	1.06	3.82	1.20
Falling Reactivity												
Standard	5.11	.87	4.92	.82	5.03	.85	5.19	.95	5.15	.79	5.09	.90
Short	5.30	.96	5.08	1.00	5.17	.97	5.29	1.07	5.28	.91	5.24	1.05
Fear												
Standard	2.48	.79	2.42	.91	2.14	.85	2.68	.88	2.02	.66	2.78	1.05
Short	2.51	.89	2.42	.99	2.16	.96	2.73	1.11	2.01	.67	2.83	1.23
High Intensity Pleasure												
Standard	5.57	.71	5.73	.80	4.96	1.00	5.92	.71	5.22	.95	5.89	.80
Short	5.60	.79	5.73	.88	4.98	1.10	5.95	.75	5.15	1.07	5.97	.85
Low Intensity Pleasure												
Standard	4.84	.87	5.27	.80	4.94	1.01	4.90	.99	5.10	.86	5.22	.92
Short	4.90	.97	5.31	.85	5.21	1.01	4.89	1.06	5.23	.86	5.23	.95
Perceptual Sensitivity												
Standard	3.75	.89	3.97	1.10	3.17	1.11	3.96	.82	3.33	1.12	4.11	1.19
Short	3.43	1.09	3.63	1.35	2.95	1.30	3.72	10.9	2.94	1.15	3.82	1.40

Sadness												
Standard	3.35	.82	3.44	.97	3.20	.90	3.49	.98	3.31	.96	3.22	.84
Short	3.55	.96	3.63	1.08	3.43	1.01	3.71	1.16	3.50	.97	3.36	.94
Smiling and Laughter												
Standard	4.52	.98	4.79	.99	4.20	1.17	4.74	1.00	4.51	1.00	4.86	1.07
Short	4.52	1.03	4.72	1.11	4.04	1.24	4.80	1.06	4.37	1.09	4.88	1.15
Soothability												
Standard	4.93	.68	4.97	.65	4.84	.72	5.12	.62	4.92	.79	5.06	.64
Short	5.43	.89	5.54	.90	5.43	.93	5.73	.69	5.46	.94	5.60	.79
Vocal Reactivity												
Standard	4.63	.87	4.65	.95	4.04	1.07	4.83	.91	4.23	1.07	4.81	1.04
Short	4.92	.93	5.00	1.01	4.28	1.12	5.23	.92	4.59	1.16	5.08	1.04
VSF PAS	4.44	.72	4.61	.86	3.53	.97	4.88	.74	3.74	.84	4.78	.96
VSF NEG	3.68	.79	3.62	.96	3.31	.82	3.88	.94	3.30	.84	3.72	.97
VSF ORC	4.71	.69	5.07	.70	5.05	.75	4.77	.72	4.95	.65	4.96	.76

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131. [b]*n* = 191. [c]*n* = 223. [d]*n* = 68. [e]*n* = 54. [f]*n* = 119.

Internal consistency

Alpha coefficients obtained for the scales of the standard and short forms are shown in Table 3. Alpha coefficients for the short form scales were approximately.07 lower, on average, than the corresponding values for standard scales. Of the 84 sample-specific short-form alphas calculated, only seven were below.70, two were below.65, and all were above.60. Activity Level demonstrated the lowest reliability for both standard and short forms, with short form alphas from three of the six samples below.70. Shortening the scales had the most adverse effect on the Fear, Cuddliness, and Sadness scales, with the latter demonstrating reliability below.70 in two data sets. As shown in Table 3, for the PAS, NEG, and ORC scales of the very short form, only one of 36 sample-specific alphas was below.70, and none were below.65.

Table 3 Internal consistency (Cronbach's alpha) of Infant Behavior Questionnaire—Revised standard, short, and very short form scales in Study 2.

Scale	**Braungart-Rieker**[a]	**Calkins**[b]	**Davis**[c]	**Gartstein**[d]	**Ross-Sheehy**[e]	**Porges**[f]	**Average Across Samples**
Activity Level							
Standard	.75	.72	.75	.82	.73	.83	.77
Short	.70	.64	.68	.74	.68	.78	.70
Approach							
Standard	.82	.84	.88	.86	.83	.91	.86
Short	.72	.81	.81	.79	.80	.86	.80
Cuddliness							
Standard	.86	.82	.83	.88	.79	.83	.84
Short	.76	.69	.75	.80	.62	.71	.72
Distress to Limitations							
Standard	.77	.81	.78	.85	.82	.83	.81
Short	.74	.75	.75	.79	.75	.79	.76
Duration of Orienting							
Standard	.82	.85	.87	.81	.83	.82	.83
Short	.76	.76	.78	.72	.73	.79	.76
Falling Reactivity							
Standard	.87	.80	.84	.88	.86	.87	.85
Short	.81	.76	.77	.82	.80	.84	.80
Fear							
Standard	.88	.90	.90	.88	.90	.91	.90

Short	.74	.77	.77	.78	.71	.79	.76
High Intensity Pleasure							
Standard	.82	.85	.85	.83	.87	.87	.85
Short	.80	.81	.82	.76	.84	.84	.81
Low Intensity Pleasure							
Standard	.84	.82	.87	.87	.86	.86	.85
Short	.75	.69	.76	.77	.71	.75	.74
Perceptual Sensitivity							
Standard	.81	.88	.89	.73	.83	.89	.84
Short	.77	.85	.85	.71	.80	.85	.81
Sadness							
Standard	.83	.87	.85	.87	.87	.83	.85
Short	.71	.75	.70	.79	.71	.67	.72
Smiling and Laughter							
Standard	.83	.81	.87	.82	.83	.84	.83
Short	.78	.79	.83	.77	.78	.80	.79
Soothability							
Standard	.82	.76	.79	.76	.88	.76	.80
Short	.81	.78	.76	.71	.85	.72	.77
Vocal Reactivity							
Standard	.81	.82	.87	.80	.87	.86	.84
Short	.77	.76	.81	.72	.82	.77	.78

VSF PAS	.70	.76	.80	.68	.74	.92	.77
VSF NEG	.73	.80	.72	.75	.79	.88	.78
VSF ORC	.73	.71	.75	.76	.75	.82	.75

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131. [b]*n* = 191. [c]*n* = 223. [d]*n* = 68. [e]*n* = 54. [f]*n* = 119.

Standard-to-short-form relations

To assess the correspondence between the standard and short scales, Levy's (1967) correction was applied. This correction removes common error variance between the two forms to achieve “true score” correlations between long scales and shorter scales extracted from the same data. As shown in Table 4, considerable consistency between the original and abbreviated scales was observed, with corrected correlation coefficients above.70 in all data sets for 12 of the 14 scales. Correspondence was relatively low for Activity Level and quite low for Soothability, for which correlations >.60 were calculated in four of six samples. Corrected standard-to-very-short correlations ranged from.71 to.86 in the individual samples.

Table 4 Corrected standard-to-short and standard-to-very short correlations of Infant Behavior Questionnaire–Revised scales in Study 2.

Scale	**Braungart-Rieker[a]**	**Calkins[b]**	**Davis[c]**	**Gartstein[d]**	**Ross-Sheehy[e]**	**Porges[f]**	**Average Across Samples**
Standard-to-short							
Activity Level	.69	.64	.70	.76	.63	.80	.70
Approach	.80	.86	.87	.85	.86	.90	.86
Cuddliness	.83	.78	.80	.82	.74	.80	.80
Distress to Limitations	.75	.77	.74	.80	.76	.80	.77
Duration of Orienting	.78	.79	.80	.77	.75	.81	.78
Falling	.83	.77	.80	.84	.83	.84	.82

Reactivity							
Fear	.84	.86	.84	.85	.80	.86	.84
High Intensity Pleasure	.80	.83	.83	.80	.86	.84	.83
Low Intensity Pleasure	.81	.78	.83	.84	.80	.81	.81
Perceptual Sensitivity	.80	.87	.87	.73	.78	.88	.82
Sadness	.78	.83	.80	.85	.76	.78	.80
Smiling and Laughter	.81	.81	.85	.79	.81	.81	.81
Soothability	.69	.62	.64	.47	.77	.59	.63
Vocal Reactivity	.79	.80	.84	.76	.87	.83	.82
Standard-to-very short							
PAS	.83	.86	.88	.71	.86	.92	.84
NEG	.82	.86	.78	.87	.89	.86	.85
ORC	.80	.77	.82	.77	.81	.79	.79

Note. PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131. [b]*n* = 195. [c]*n* = 223. [d]*n* = 72. [e]*n* = 54. [f]*n* = 119.

Longitudinal stability and estimated retest reliability

Pearson's correlations were calculated to assess longitudinal stability of the standard, short, and very short scales across all time points in the Braungart-Reiker data set. As shown in Table 5, stability coefficients for the short form were very similar to those obtained with standard form scales. When averaged across all time spans and scales, stability coefficients averaged.43 for standard scales and.42 for short form scales. Decreases in stability from the standard to short form were most substantial for the Cuddliness and Duration of Orienting scales, for which the correlations decreased by >.06 for two age spans. In contrast, stability was higher for the short Vocal Reactivity scale, relative to the long form version, across several time spans.

Table 5 Longitudinal stability, and estimated retest reliability of Infant Behavior Questionnaire–Revised standard, short, and very short form scales in Study 2 (Braungart-Reiker sample).

	Longitudinal Stability Time Span						
Scale	**2 Months**	**4 Months**	**5 Months**	**7 Months**	**9 Months**	**11 Months**	**Estimated Retest Reliability**
Activity Level							
Standard	.53	.43	.51	.35	.29	.26	.68
Short	.48	.39	.51	.32	.27	.24	.63
Approach							
Standard	.45	.35	.48	.21	.17	.23	.58
Short	.43	.38	.48	.19	.17	.26	.55
Cuddliness							
Standard	.57	.52	.51	.47	.47	.37	.65
Short	.54	.44	.51	.46	.46	.28	.66
Distress to Limitations							
Standard	.51	.32	.53	.45	.41	.22	.67
Short	.52	.34	.53	.46	.43	.26	.65
Duration of Orienting							
Standard	.50	.32	.51	.49	.24	.12	.89
Short	.46	.26	.51	.46	.21	.06	.89
Falling Reactivity							
Standard	.67	.56	.54	.56	.46	.45	.77
Short	.63	.49	.54	.57	.46	.44	.73

Fear							
Standard	.57	.37	.52	.49	.32	.16	.88
Short	.55	.34	.52	.47	.32	.14	.82
High Intensity Pleasure							
Standard	.48	.51	.47	.35	.26	.15	.74
Short	.48	.49	.47	.34	.28	.19	.65
Low Intensity Pleasure							
Standard	.53	.36	.67	.42	.21	.15	.90
Short	.54	.37	.67	.41	.20	.14	.93
Perceptual Sensitivity							
Standard	.53	.50	.63	.36	.35	.32	.60
Short	.51	.42	.63	.36	.33	.36	.57
Sadness							
Standard	.57	.50	.62	.52	.45	.39	.69
Short	.56	.47	.62	.50	.44	.32	.70
Smiling and Laughter							
Standard	.63	.54	.67	.56	.43	.29	.86
Short	.62	.49	.67	.55	.38	.24	.90
Soothability							
Standard	.48	.36	.48	.37	.35	.34	.55
Short	.48	.33	.48	.37	.32	.37	.54
Vocal Reactivity							

Standard	.63	.47	.53	.52	.37	.30	.81
Short	.64	.51	.53	.51	.42	.39	.74
VSF PAS	.59	.56	.44	.41	.36	.40	.64
VSF NEG	.59	.46	.60	.46	.33	.19	.88
VSF ORC	.52	.44	.59	.52	.41	.26	.70

Note. 2-month span values are the average of correlations from 3 to 5 months (*n* = 126), 5 to 7 months (*n* = 114), and 12 to 14 months (*n* = 102). 4-month span is from 3 to 7 months (*n* = 115). 5-month span is from 7 to 12 months (*n* = 109). 7-month span is average of 5 to 12 months (*n* = 114) and 7 to 14 months (*n* = 101). 9-month span is average of 3 to 12 months (*n* = 114) and 5 to 14 months (*n* = 101). 11-month span is from 3 to 14 months (*n* = 105). Estimated retest reliability was derived using a formula devised by Heise (1969) for stability correlations obtained at three time points. The coefficients reported are the average of values obtained over all possible three-time-point combinations. PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.

The longitudinal correlations across 2- to 12-month time spans confound two factors: the retest reliability of the scores and the true stability of the underlying traits. Heise (1969) argued that these can be disentangled when three time points are observed, such that retest reliability can be estimated as $(r_{12} * r_{23}) / r_{13}$. The final column in Table 6 contains estimates of retest reliability obtained with this formula. Specifically, these values are the average retest reliability estimates over the 10 possible combinations of three data points. The average retest estimates were nearly identical for the short and standard scales of the IBQ–R, with the exception of High Intensity Pleasure and Vocal Reactivity, for which the short form reliabilities were.09 and.07 lower than the standard form. The levels of reliability obtained with both short and standardized scales are comparable to those obtained, using Heise's formula, by Terracciano, Costa, and McCrae (2006) for scales of the Revised NEO Personality Inventory (NEO–PI–R).

Table 6 Means, standard deviations, and internal consistency (Cronbach's alpha) of Infant Behavior Questionnaire–Revised short and very short form scales in Study 3.

	Brand[a]			**Sullivan**[b]			**Leerkes**[c]			**Horodynski**[d]		
Scale	***M***	***SD***	**α**	***M***	***SD***	**α**	***M***	***SD***	**α**	***M***	***SD***	**α**
Activity Level	4.22	.94	.68	3.96	.96	.69						
Approach	5.90	.73	.79	4.92	1.00	.77						
Cuddliness	5.40	.86	.81	6.02	.70	.69						

Distress to Limitations	4.05	1.24	.83	3.81	.92	.67	3.70	.99	.75			
Duration of Orienting	4.16	1.00	.69	4.41	1.14	.82						
Falling Reactivity	5.66	.99	.86	5.22	.82	.74	5.13	.97	.76			
Fear	3.23	1.08	.74	2.35	.88	.78	2.61	1.13	.83			
High Intensity Pleasure	6.34	.66	.78	5.91	.77	.76						
Low Intensity Pleasure	5.46	.87	.70	5.54	.73	.66						
Perceptual Sensitivity	4.59	1.20	.80	3.56	1.37	.87						
Sadness	3.52	1.03	.74	3.56	.91	.71	3.37	.94	.70			
Smiling and Laughter	5.28	.91	.75	4.77	1.08	.79						
Soothability	5.84	.75	.84	5.54	.84	.79	5.53	.92	.77			
Vocal Reactivity	5.68	.83	.76	5.02	1.00	.79						
VSF PAS							5.06	.78	.76	3.50	1.12	.80
VSF NEG							3.46	.91	.79	3.43	1.10	.81
VSF ORC							5.47	.63	.71	5.26	.75	.74

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 54. [b]*n* = 86. [c]*n* = 224. [d]*n* = 225.

Discussion

The results of Study 2 suggest that the shortened versions of the IBQ–R scales demonstrate an adequate degree of reliability without unduly compromising the content validity of the instrument. The 14 scales of the IBQ–R short form were tested in six independent data sets that collectively represent all ages for which the IBQ–R was designed, and the vast majority of alphas generated in these analyses were greater than.70. This degree of internal consistency compares favorably to other instruments: In a recent review of temperament questionnaire methodology, of 43 instruments summarized, 28 contained at least one scale with an internal consistency estimate lower than.70, and 15 included at least one scale with a reported internal consistency less than.60 (Gartstein et al., 2012). Thus, all of the scales making up the IBQ–R short form demonstrate coherence to a degree considered acceptable in the temperament literature. Because the Activity Level and Cuddliness scales exhibited reliability under.70 for more than one sample, researchers with a particular interest in these specific attributes might wish to consider using the longer form in their investigations.

It is also the case that, when making decisions regarding retention of items, internal consistency was only one of the criteria employed. Several items that demonstrated high correlations with others in their scale were omitted and replaced, so that the short scales more fully represented the content of the standard IBQ–R scales. The corrected long–short correlations quantify the degree of correspondence between the standard and abbreviated forms of the scales. For most scales, these values were similar in magnitude to those obtained in other short form investigations (e.g., Petrides, Jackson, Furnham, & Levine, 2003). In some data sets, however, the Soothability score exhibited low long-to-short-form correlations, likely as a function of the item selection process for this scale. As noted in regard to scale construction, factor analyses of the standard Soothability scale revealed four factors, two of which were not retained for the short form of the IBQ–R. Although this resulted in a short Soothability scale with higher alphas than the standard form scale, the enhanced internal consistency came at the cost of content validity, and interpretations of the short scale should take into account these limitations regarding the specificity of infants' soothing tendencies. Differences in the content of the short and standard forms of the Cuddliness scale might also explain the substantially lower stability coefficients obtained with the short version. For this scale, items regarding infant reactions when the caregiver returned from an absence were omitted in forming the short versions. It is possible that positive emotionality during such episodes is particularly stable during the first year, in comparison to positivity during moments of play or caregiving, which are the contexts described in the items retained for the short form. Overall, the stability of the standard, short, and very short scales of the IBQ–R is comparable to the levels of continuity obtained over similar intervals by individual scales from a number of different infant temperament measures (see review by Slabach et al., 1991), as are the estimates of retest reliability (see review by Gartstein, 2012).

Study 3: Samples Administered the Short and Very Short Forms

Study 3 addresses the possibility that, in comparison with analyses conducted on standard form data, the psychometric properties of the short and very short form scales might be compromised when the abbreviated forms themselves are administered, or when a hybrid of the short and very short forms is used. Study 3 also affords an opportunity to explore a related psychometric issue, namely, whether or not the internal consistency of our measures is robust to deviations from the prescribed age range (i.e., when the short versions of the IBQ–R are administered to parents of children older or younger than 3–12 months of age). Finally, because the investigators for these studies shared data regarding the race and income level of their participants, and because these samples were quite diverse, we were able to assess the impact of these variables on internal consistency, calculating internal consistency for White and African American participants, and those above and below the poverty line.

Samples

Two short form data sets were acquired. The first was gathered by Rebecca Brand of Villanova and included 54 infants (13 female, 18 male, 23 missing gender data) ranging in age from 20 to 102 weeks (*M* = 51.5, *SD* = 17.6). Of the 34 families providing demographic data, 28 infants were White, one was African American, and five were multiracial; mothers ages ranged from 23 to 43 (*M* = 32); all mothers were high school graduates, and 28 had also completed college. The second was gathered by Margaret Sullivan of the University of Medicine and Dentistry, New Jersey, and included 86 infants (42 female) ranging from 4.5 to 6 months of age (*M* = 5.02, *SD* =.26). Fifty-seven percent of these infants were White, 7% were African or African American, 8% were Hispanic, 7% were Asian/Indian, and 20% were multiracial. One set of very short form data was collected by Mildred Horodynski of Michigan State University, and included scores for 225 (100 female) infants ranging from birth to 5 months of age (*M* = 1.83, *SD* = 1.07). Of the 218 reporting ethnicity, 126 mothers were African American, 43 were White, 3 were Asian, and 42 were multiracial. Of the 199 reporting household income, 122 reported a household income under $10,000, with 28 reporting between $10,000 and $15,000, 25 between $15,000 and $25,000, and 24 over $25,000. Finally, Esther Leerkes of University of North Carolina at Greensboro contributed data collected using a hybrid instrument consisting of the 37 very short form items, as well as the remaining items from the short form Distress to Limitations, Fearfulness, Falling Reactivity, and Soothability scales. These data contained reports from mothers of 224 (115 female; 93 White, 96 African American, 25 more than one race, and 10 Hispanic or other race) 6-month-old infants, 87 of whom were residing in families in poverty, defined as an income-to-needs ratio below two.

Results

Descriptive statistics and Cronbach's alphas for the short form scales are shown in Table 6. In the Brand data set, all scales except Activity Level (α =.68), demonstrated alphas of.70 or greater. In the Sullivan data set, 10 alphas were greater than.70 and none were below.65. In the Horodynski very short form data set, all alphas were greater than.70, as were all alphas for short and very short scales included in Leerkes's “hybrid” form.

To assess potential differences in internal consistency that might be associated with respondent race, the Horodynski and Leerkes samples were analyzed separately for African Americans and Whites. In the Horodynski sample, alphas for PAS, NEG, and ORC were.77,.83, and.75 for African American subjects and.85,.81, and.78 for the White subsample. In the Leerkes sample, the corresponding values were.69,.78, and.72 for the African American subsample, and.77,.78, and.69 in the White subsample. Alphas for the Distress to Limitations, Falling Reactivity, Fear, Sadness, and Soothability scales were.72,.71,.79,.71, and.75 for African American respondents, and.76,.78,.82,.70, and.82 for Whites.

We also investigated household income as a potential influence on internal consistency. In the Horodynski data set, alphas were calculated separately for participants reporting household incomes more or less than $10,000. Alphas for PAS, NEG, and ORC were.78,.78, and.73 in the former subsample, and.81,.78, and.75 in the latter. In the Leerkes sample, these values were calculated for those with income-to-needs ratio above and below two. Alphas for PAS, NEG, and ORC were.76,.80, and.71 in the former, and.75,.79, and.72 in the latter. Alphas for Distress to Limitations, Falling Reactivity, Fear, Sadness, and Soothability were.79,.77,.83,.66, and.78 in the former and.69,.75,.82,.75, and.76 in the latter.

Discussion

When administered to two independent samples, the IBQ–R short form generated internal consistency estimates similar to those obtained when item scores had been extracted from data collected with the standard forms. It is notable that this degree of reliability was evident even in a sample for which nearly half of the infants were over 1 year, the oldest age investigated in developing the standard IBQ–R (Gartstein & Rothbart, 2003). This suggests the usefulness of the short form for researchers who wish to include children up to 2 years of age in their longitudinal or cross-sectional investigations.

The three scales of the very short form demonstrated adequate internal consistency, although a number of the infants who received parental ratings were well under the age of 3 months. The original IBQ (Rothbart, 1981) has occasionally been used successfully with neonates (Worobey, 1986), but to our knowledge, no published studies of very young infants have relied on the standard IBQ–R to date. Additional research is required to indicate whether the various scales of the standard and short form instruments are appropriate for very young infants, but our findings suggest that the three broad factors of the very short form can be measured coherently in such a sample. In addition, the internal consistency of the very short form scales, and select short form scales, was similar in White and African American subsamples, and for mothers from both middle and lower economic ranges, confirming the usefulness of the instrument across a wide range of potential subjects. Finally, favorable psychometric characteristics of the very short form were preserved even when administered in a manner that overemphasized negative affectivity dimensions. This innovative strategy allows researchers flexibility to examine both broad and narrow aspects of child behavior when such an approach is warranted by their specific research questions.

Study 4: Convergent and Predictive Validity of Select Scales

Two published reports have investigated the correspondence between standard IBQ–R scales and observational data obtained through structured laboratory episodes. The authors of these papers agreed to replicate the relevant analyses using short form scales extracted from their data. Gartstein and Marmion (2008; see Study 2 for sample characteristics) assessed infant fear across two procedures (exposure to a stranger and to masks) and derived a single variable representing facial, body, and vocal fear, reporting a correlation of.28 (*p* <.05) between this variable and the standard IBQ–R Fear scale. When replicated with the short form scale, this correlation was the same, *r*(65) =.28, *p* <.05. Gartstein and Marmion (2008) also assessed positive affect during peek-a-boo and free play contexts with the mother, reporting a nonsignificant correlation of.20 (*ns*) between this variable and the standard IBQ–R Smiling and Laughter scale. The short form scale also failed to correlate significantly with observed positive affect,*r*(65) =.14, *ns*. Parade and Leerkes (2008; see Study 1 for sample characteristics) assessed infant fear and anger via composites of affective and motor indexes during exposure to a loud and unfamiliar toy (fear) and arm restraint (anger), correlating these with maternal and paternal ratings of IBQ–R Approach, Fear, and Distress to Limitations. The observational fear score was positively correlated with maternal ratings of IBQ–R Fear, *r*(98) =.22, *p* <.05, and negatively correlated with paternal ratings of IBQ–R Approach, *r*(66) = –.28, *p* <.05. When calculated with short form scales, the corresponding correlations were *r*(98) =.17, *p* <.10, and *r*(66) = –.30, *p* <.05, respectively. Observed anger was correlated with mothers' standard IBQ–R fear ratings, *r*(98) =.32, *p* <.01. The corresponding correlation using the short form scale was also significant, *r*(98) =.24, *p* <.05. Nonsignificant correlations reported by Parade and Leerkes (2008) remained nonsignificant when calculated with short form scales.

Putnam et al. (2008) also published data relevant to the validity of the IBQ–R, exploring longitudinal relations between standard form IBQ–R data (measured between 3 and 12 months child age) and corresponding scales on the Early Childhood Behavior Questionnaire (ECBQ; Putnam, Gartstein, & Rothbart, 2006; measured between 18 and 32 months) and CBQ (Rothbart et al., 2001; measured between 37 and 59 months). This sample was primarily White, with an average family income of $41,798 (*SD* = $19,154), average maternal age of 31 (*SD* = 5.3), and mothers had an average of 14.5 (*SD* = 2.4) years of education. In the analyses reported by Putnam et al. (2008), significant correlations for scales addressing 11 constructs measured with both the IBQ–R and ECBQ were significant, demonstrating relative longitudinal consistency. Specifically, stability coefficients from the IBQ–R to the toddler measure for High Intensity Pleasure, Activity Level, Approach, Perceptual Sensitivity, Frustration, Sadness, Falling Reactivity, Fear, Duration of Orienting, Low Intensity Pleasure, and Cuddliness, *r*s(248) =.30,.32,.32,.45,.22,.24,.30, 23,.23,.34, and.36, respectively, *p*s <.01, (average *r* =.30). When these analyses were repeated using short form IBQ–R scales, all correlations were significant, *r*s(248) =.27,.32,.28,.34,.17,.22,.23,.19,.24,.34, and.31, *p*s <.01, although somewhat smaller than those obtained with standard scales (average *r* =.26). Six of the 11 scales measuring constructs on both the IBQ–R and CBQ were relatively stable longitudinally: Activity Level, Approach, Smiling and Laughter, Frustration, Sadness, and Perceptual Sensitivity, *r*s(140) =.22,.23,.26,.29,.18,.23, respectively,*p*s <.05. When calculated with short form IBQ–R scales, correlations for the first five of these remained significant,*r*s(140) =.20,.30,.30,.25,.19, *p*s <.05, although the correlation for Perceptual Sensitivity was not, *r*(140) =.10, *ns*.

Discussion

Because both observational and questionnaire methods are prone to error (Rothbart & Bates, 2006), it is reassuring that a degree of correspondence exists between infant behavior during brief laboratory tasks and the relevant scales of the standard and short IBQ–R. Although all convergent validity correlations shown to be significant with the standard form scales were at least marginally significant when short form scales were employed, many were lower in magnitude, suggesting that researchers desiring the greatest level of convergence among multimethod measures should continue to use the standard form when possible. By the same token, researchers are urged to expand their observational methodology to involve multiple tasks in their assessment of individual differences of infants' reactivity and regulation. Similarly, the short form scales demonstrated statistically significant prediction to parent-rated temperament at older ages at levels that were slightly below that of standard IBQ–R scales, suggesting that some relevant information was not retained when fewer items were used to gauge infant temperament. Despite these caveats, the results of Study 4 suggest that the short form retains a great deal of the validity evident in the longer instrument.

General Discussion

The results of our analyses suggest that the short and very short forms of the IBQ–R are valuable tools for researchers who wish to incorporate temperament into their protocols, but are hesitant to administer the original 191-item instrument due to time demands on subjects. The abbreviated scales of the short form are strongly correlated with, exhibit levels of interparent agreement that are nearly identical to, and demonstrate only slightly lower levels of internal consistency, longitudinal stability, and convergence with observational data than their long versions. In addition, the very short form assesses three broad, empirically derived, and conceptually meaningful traits with levels of reliability and stability that are similar to those obtained with the more discrete scales comprising the IBQ–R and other temperament measures (Gartstein, 2012). Because reliability and validity are not always consistent when measures are used across different samples (Thompson, 1994), the use of multiple samples represents a substantial strength of this investigation. In particular, the demonstration of acceptable qualities in samples that were economically and racially diverse, and involved children younger than 3 months of age and up to the age of 2 years, yield confidence that the measures could be successfully employed in research including a variety of populations.

It is anticipated that the very short form will be of most use for large-scale investigations, including epidemiological studies, for which a wide variety of constructs are assessed. It is estimated that the 191-item standard IBQ–R takes parents approximately 1 hour to complete, suggesting that the 37-item very short form can be completed by most parents in under 12 minutes. The short form, at 91 items, can be filled out in approximately 30 minutes, and thus would be appropriate for developmental scholars who wish to assess a wide variety of temperament attributes, but are relatively constrained with respect to the demands they can place on parent participants. In addition, investigators with very specific research questions and severe limits on subject demand might choose to administer only a select number of short scales to assess the discrete traits in which they are particularly interested. Consideration should also be given to the use of “hybrid” measures, such as those utilized by Leerkes in the data she contributed to Study 3, when appropriate given the goals and hypotheses of a particular investigation. The strategic combination of extensive measures of fine-grained traits (e.g., multiple types of negative affectivity) to match narrow research questions, combined with efficient assessment of other broad factors to facilitate more exploratory analyses, allows for flexible pursuit of both inductive and deductive knowledge regarding the correlates of temperament, which might be desirable for certain research designs. Researchers with interests in discrete traits are advised to carefully consider the implications of abbreviation for individual scales when deciding on whether to use the standard or short versions.

There are limitations to our investigation. The fundamental goal of these studies was to evaluate the short forms with respect to their longer counterparts, which were developed largely through a rational approach to scale construction. Given this focus, our analyses do not provide a substantial advance in the understanding of the structure of temperament. With respect to this issue, it is worth noting that the three-factor structure that initially emerged from exploratory factor analyses of the standard form by Gartstein and Rothbart (2003) required alterations on the basis of modification indexes to achieve good fit in subsequent studies (Gartstein et al., 2005; Montirosso, Cozzi, Putnam, Gartstein, & Borgatti, 2011). Confirmatory factor analyses of Study 2 data suggested good fit (e.g., Comparative Fit Index =.955, root mean square error of approximation =.058) of short form data to a model derived from standard form scores, suggesting that interrelations among the scales is similar across the two versions of the IBQ–R. However, more precise exploration of relations of items across scales of the IBQ–R represents a valuable direction for future investigations, possibly revealing latent factors that are not apparent in scale-level analyses.

Additional shortcomings are based in the nature of the data to which we were granted access. Agreement between the short and standard forms was not assessed directly by administering both to the same sample. Although we statistically controlled for shared error when calculating standard–short-form correlations from data collected with a standard form, a more direct comparison would lead to greater confidence in the correspondence between the two versions of the IBQ–R. Another limitation is the relative lack of diversity in the majority of the included samples. Although racial and socioeconomic diversity were represented to a degree in Study 3, no samples were drawn from studies of clinical populations, or from populations for which English was a second language. The standard IBQ–R has recently been used to examine cross-cultural differences and similarities in temperament (e.g., Gartstein et al., 2005; Montirosso et al., 2011), but it is not yet known whether the items chosen for the shortened versions will be similarly useful when used with respondents from other cultures. Predictive validity was not examined for all scales, and data concerning convergent validity with observational methods was available for only a few scales. In addition, our assessments of external validity were carried out with short form scale scores that were extracted from standard form data, and it is possible that different findings might be obtained when the short form itself is administered. Future studies investigating these forms of validity in not only the short and very short forms of the IBQ–R, but the standard form as well, are necessary to confirm the convergence of these instruments with other assessment methods.

Acknowledgments

The data used in this effort were collected through the support of USDA grant 2009-55215-05220; NIMH grants 1F32-HD055040-01, R01-HD049878, R03-MH61831A, R03-MH0670, R01-MH061788, R03-HD048691, R01-HD 053570-01, R03-HD39802, R03-HD043057, R01-HD049878, R01-HD52669, R01-HD51852; and Villanova University. Preliminary findings were presented at the Biennial Meeting of the Society for Research in Child Development in Denver, CO, April 2009. We wish to express our gratitude to the researchers (indicated in the text) who generously donated their IBQ-R data to this effort, as well as their collaborators, especially Laurie Van Egeren, Keri Heilman, and Erica Bullock. Copies of the standard, short, and very short forms of the IBQ-R can be acquired from http://www.bowdoin.edu/∼sputnam/rothbart-temperament-questionnaires/.

Notes

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 68. [b]*n* = 64. [c]*n* = 55. [d]*n* = 43. [e]*n* = 49. [f]*n* = 79.

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131. [b]*n* = 191. [c]*n* = 223. [d]*n* = 68. [e]*n* = 54. [f]*n* = 119.

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131. [b]*n* = 191. [c]*n* = 223. [d]*n* = 68. [e]*n* = 54. [f]*n* = 119.

Note. PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 131.[b]*n* = 195. [c]*n* = 223. [d]*n* = 72. [e]*n* = 54. [f]*n* = 119.

Note. 2-month span values are the average of correlations from 3 to 5 months (*n* = 126), 5 to 7 months (*n* = 114), and 12 to 14 months (*n* = 102). 4-month span is from 3 to 7 months (*n* = 115). 5-month span is from 7 to 12 months (*n* = 109). 7-month span is average of 5 to 12 months (*n* = 114) and 7 to 14 months (*n* = 101). 9-month span is average of 3 to 12 months (*n* = 114) and 5 to 14 months (*n* = 101). 11-month span is from 3 to 14 months (*n* = 105). Estimated retest reliability was derived using a formula devised by Heise (1969) for stability correlations obtained at three time points. The coefficients reported are the average of values obtained over all possible three-time-point combinations. PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.

Note. VSF = very short form; PAS = Positive Affectivity/Surgency; NEG = Negative Emotionality; ORC = Orienting/Regulatory Capacity.[a]*n* = 54. [b]*n* = 86. [c]*n* = 224. [d]*n* = 225.

References

1. Braungart-Rieker, J.M., Hill-Soderlund, A.L., & Karrass, J. (2010). Fear and anger reactivity trajectories from 4 to 16 months: The roles of temperament, regulation, and maternal sensitivity. Developmental Psychology, 46, 791–804.

2. Buss, A.H., & Plomin, R. (1984). Temperament: Early developing personality traits. Hillsdale, NJ: Erlbaum.

3. Canivez, G.L., & Watkins, M.W. (2010a). Exploratory and higher-order factor analyses of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV) adolescent subsample. School Psychology Quarterly, 25, 223–235.

4. Canivez, G.L., & Watkins, M.W. (2010b). Investigation of the factor structure of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. Psychological Assessment, 22, 827–836.

5. Carey, W.B., & McDevitt, S.C. (1978). Revision of the infant temperament questionnaire. Pediatrics, 61, 735–739.

6. De Lauzon-Guillain, B., Wijndaele, K., Clark, M., Acerini, C.L., Hughes, I.A., Dunger, D.B., … Ong, K.K. (2012). Breastfeeding and infant temperament at age three months. PLoS ONE, 7, ArtID e29326.

7. Enders, C. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. Educational and Psychological Measurement, 64, 419–436.

8. Evans, D., & Rothbart, M.K. (2007). Developing a model for adult temperament. Journal of Research in Personality, 41, 868–888.

9. Gartstein, M.A. (2012). Asking questions about temperament: self- and other-report measures across the lifespan. In M. Zentner & R. Shiner (Eds.), Handbook of temperament (pp. 183–208). New York, NY: Guilford.

10. Gartstein, M.A., Bridgett, D.J., Rothbart, M.K., Robertson, C., Iddins, E., Ramsay, K., & Schlect, S. (2010). A latent growth examination of fear development in infancy: Contributions of maternal depression and the risk for toddler anxiety. Developmental Psychology, 46, 651–658.

11. Gartstein, M.A., Knyazev, G.G., & Slobodskaya, H.R. (2005). Cross-cultural differences in the structure of infant temperament: United States of America (U.S.) and Russia. Infant Behavior & Development, 28, 54–61.

12. Gartstein, M.A., & Marmion, J. (2008). Fear and positive affectivity in infancy: Convergence/discrepancy between parent-report and laboratory-based indicators. Infant Behavior & Development, 31, 227–238.

13. Gartstein, M.A., Putnam, S.P., & Rothbart, M.K. (2012). Etiology of preschool behavior problems: Contributions of temperament attributes in early childhood. Infant Mental Health Journal, 33, 197–211.

14. Gartstein, M.A., & Rothbart, M.K. (2003). Studying infant temperament via the revised infant behavior questionnaire. Infant Behavior and Development, 26, 64–86. **15.** George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update *(4th ed.)*. Boston, MA: Allyn & Bacon.

16. Goodwin, L.D., & Goodwin, W.L. (1999). Measurement myths and misconceptions. School Psychology Quarterly, 14, 408–427.

17. Heise, D.R. (1969). Separating reliability and stability in test–retest correlation. American Sociological Review, 34, 93–101.

18. Kagan, J., Reznick, J.S., & Snidman, N. (1987). The physiology and psychology of behavioral inhibition in children. Child Development, 58, 1459–1473.

19. Knapp, T.R., & Brown, J.K. (1995). Ten measurement commandments that often should be broken. Research on Nursing & Health, 18, 465–469.

20. Kochanska, G., Coy, K.C., Tjebkes, T.L. & Husarek, S.J. (1998). Individual differences in emotionality in infancy. Child Development, 64, 375–390.

21. Kopalle, P.K., & Lehmann, D.R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. Organizational Behavior and Human Decision Processes, 70, 189–197.

22. Levy, P. (1967). The correction for spurious correlation in the evaluation of short-form tests. Journal of Clinical Psychology, 23, 84–86.

23. Loevinger, J. (1954). The attenuation paradox in test theory. Psychological Bulletin, 51, 493–504.

24. Montirosso, R., Cozzi, P., Putnam, S.P., Gartstein, M.A., & Borgatti, R. (2011). Studying cross-cultural differences in temperament in the first year of life: United States and Italy. International Journal of Behavioral Development, 35, 27–37.

25. Morasch, K.C., & Bell, M.A. (2012). Self-regulation of negative affect at 5 and 10 months. Developmental Psychobiology, 54, 215–221.

26. Nunnally, J.C. (1978). Psychometric theory *(2nd ed.)*. New York, NY: McGraw-Hill.

27. Parade, S.H., & Leerkes, E.M. (2008). The reliability and validity of the Infant Behavior Questionnaire–Revised. Infant Behavior and Development, 31, 637–646.

28. Petrides, K.V., Jackson, C.J., Furnham, A., & Levine, S.Z. (2003). Exploring issues of personality measurement and structure through the development of a short form of the Eysenck personality profiler. Journal of Personality Assessment, 81, 271–280.

29. Putnam, S.P., Ellis, L.K., & Rothbart, M.K. (2001). The structure of temperament from infancy through adolescence. In A. Eliasz & A. Angleitner (Eds.), Advances/proceedings in research on temperament (pp. 165–182). Lengerich, Germany: Pabst Scientist.

30. Putnam, S.P., Gartstein, M.A., & Rothbart, M.K. (2006). Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. Infant Behavior and Development, 29, 386–401.

31. Putnam, S.P., & Rothbart, M.K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. Journal of Personality Assessment, 87, 102–112.

32. Putnam, S.P., Rothbart, M.K., & Gartstein, M.A. (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. Infant and Child Development, 17, 387–405.

33. Rothbart, M.K. (1981). Measurement of temperament in infancy. Child Development, 52, 569–578.

34. Rothbart, M.K. (1986). Longitudinal observation of infant temperament. Developmental Psychology, 22, 356–365.

35. Rothbart, M.K. (2011). *Becoming who we are: Temperament and personality in development*. New York, NY: Guilford.

36. Rothbart, M.K., Ahadi, S.A., & Evans, D.E. (2000). Temperament and personality: Origins and outcomes. Journal of Personality and Social Psychology, 78, 122–135.

37. Rothbart, M.K., Ahadi, S.A., Hershey, K.L., & Fisher, P. (2001). Investigations of temperament at 3–7 years: The Children's Behavior Questionnaire. Child Development, 72, 1394–1408.

38. Rothbart, M.K., & Bates, J.E. (2006).Temperament. In N. Eisenberg & W. Damon (Eds.) Handbook of child psychology: Vol. 3. Social, emotional, and personality development *(6th ed*., pp. 99–166). New York, NY: Wiley.

39. Rothbart, M.K., & Mauro, J.A. (1990). Questionnaire approaches to the study of infant temperament. In J.W. Fagen & J. Colombo (Eds.), Individual differences in infancy: Reliability, stability and prediction (pp. 411–429). Hillsdale, NJ: Erlbaum.

40. Slabach, E.H., Morrow, J., & Wachs, T.D. (1991). Questionnaire measurement of infant and child temperament: Current status and future directions. In J. Strelau & A. Angleitner (Eds.), Explorations in temperament: International perspectives on theory and measurement (pp. 337–358). New York, NY: Plenum.

41. Smith, G.T., McCarthy, D.M., & Anderson, K.G. (2000). On the sins of short-form development. Psychological Assessment, 12, 102–111.

42. Streiner, D.L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. Journal of Personality Assessment, 80, 99–103.

43. Terracciano, A., Costa, P.T., & McCrae, R.R. (2006). Personality plasticity after age 30. Personality and Social Psychology Bulletin, 32, 999–1009.

44. Thomas, A., & Chess, S. (1977). Temperament and development. Oxford, UK: Brunner/Mazel.

45. Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837–847.

46. Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. Measurement and Evaluation in Counseling and Development, 44, 159–168.

47. Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594–604.

48. Worobey, J. (1986). Convergence among assessments of temperament in the first month. Child Development, 57, 47–55.

49. Zentner, M., & Shiner, R. (2012). Handbook of temperament. New York, NY: Guilford.

50. Zoski, K.W., & Jurs, S. (1996). An objective counterpart to the visual scree test for factor analysis: The standard error scree. Educational and Psychological Measurement, 56, 443–451.