JOURNAL OF PERSONALITY ASSESSMENT, *87*(1), 102–112

Development of Short and Very Short Forms of the Children's Behavior Questionnaire

Samuel P. Putnam

Department of Psychology
Bowdoin College

Mary K. Rothbart

Department of Psychology
University of Oregon

Using data from 468 parents and taking into account internal consistency, breadth of item content, within-scale factor analysis, and patterns of missing data, we developed short (94 items, 15 scales) and very short (36 items, 3 broad scales) forms of the Children's Behavior Questionnaire (CBQ; Rothbart, Ahadi, & Hershey, 1994; Rothbart, Ahadi, Hershey, & Fisher, 2001), a well-established parent-report measure of temperament for children aged 3 to 8 years. We subsequently evaluated the forms with data from 1,189 participants. In mid/high-income and White samples, the CBQ short and very short forms demonstrated both satisfactory internal consistency and criterion validity, and exhibited longitudinal stability and cross-informant agreement comparable to that of the standard CBQ. Internal consistency was somewhat lower among African American and low-income samples for some scales. Very short form scales demonstrated acceptable internal consistency for all samples, and confirmatory factor analyses indicated marginal fit of the very short form items to a three-factor model.

Over the past decade, researchers have become increasingly interested in relations between individual differences in children's temperament and other important social-emotional variables including empathy, attachment, conscience, and problems in social adjustment (e.g., Guthrie et al., 1997; Kochanska, 1997; Lengua, Wolchik, Sandler, & West, 2000). This interest has resulted in a search for more efficient instruments. To help provide a response to this search, we have undertaken work to develop short and very short forms of a parent-report measure of temperament, the Children's Behavior Questionnaire (CBQ; Rothbart, Ahadi, & Hershey, 1994; Rothbart, Ahadi, Hershey, & Fisher, 2001).

The CBQ was developed to provide a highly differentiated caregiver report assessment of temperament in children 3 to 8 years of age. The instrument is grounded in a definition of temperament as constitutionally based individual differences in reactivity and self-regulation, influenced over time by heredity and experience (Rothbart & Derryberry, 1981). Domains included in the instrument include positive and negative emotion, motivation, activity level, and attention. Specific dimensions chosen for the CBQ were based on constructs of temperament in infancy, as measured by the Infant Behavior Questionnaire (Rothbart, 1981), and in adulthood as measured by the Physiological Reactions Questionnaire (Derryberry & Rothbart, 1988), and items were rationally generated based on conceptual definitions for each scale.

In the CBQ, parents are asked to rate their child on a 7-point scale ranging from 1 (*extremely untrue of your child*) to 7 (*extremely true of your child*). Parents are also provided with a *not applicable* response option when the child has not been observed in the situation described. The standard form of the CBQ consists of 195 items assessing the following 15 scales of 12 to 14 items each: Activity Level, Anger/Frustration, Approach/Positive Anticipation, Attentional Control, Discomfort, Falling Reactivity/Soothability, Fear, High Intensity Pleasure, Impulsivity, Inhibitory Control, Low Intensity Pleasure, Perceptual Sensitivity, Sadness, Smiling and Laughter, and Shyness. Scale scores are created by averaging applicable item scores.

Validation of the CBQ has been offered via a number of investigations over the past decade. The standard form has been used to study genetic and environmental influences on temperament (Goldsmith, Buss, & Lemery, 1997), longitudinal change and consistency in temperament (Murphy, Eisenberg, Fabes, Shepard, & Guthrie, 1999; Tomlinson, Harbaugh, & Anderson, 1996) as well as cross-cultural similarities and differences in the structure of temperament (Ahadi, Rothbart, & Ye, 1993). In addition, both the overall instrument and select scales have been employed in studies of temperament in relation to a variety of topics including perceived competence (Schaughency & Fagot, 1993), temperamental types or clusters in preschoolers (Aksan et al., 1999), ability estimation and injury proneness (Schwebel & Plumert, 1999), problem behaviors (Eisenberg, Fabes, Guthrie, & Murphy, 1996; Lengua, West, & Sandler, 1998), mental development and the ability to delay gratification (Silverman & Ippolito, 1995), prosocial behavior (Eisenberg, Fabes, Karbon, & Murphy, 1996), mothers' perceptions of power and patterns of control (Mills, 1998), social competence in peer interactions (Fabes et al., 1999), parents' reactions to children's negative emotions (Eisenberg et al., 1999), and physiological stress responses such as cortisol production and cardiac vagal tone (Donzella, Gunnar, Krueger, & Alwin, 2000). Provision of a short form of the instrument may benefit researchers who wish to include a fine-grained temperament measure in a multivariate investigation but for whom space limitations make the standard form of the CBQ inappropriate. For researchers who are severely restricted with respect to participant resources, the very short form will allow for efficient measurement of three empirically derived and theoretically informative broad aspects of temperament.

A number of considerations guided the construction of the short and very short forms. As in the development of other measures, we sought to maximize the reliability and validity of these instruments. Reliability and content validity, however, are often in conflict during the construction of short forms. When questionnaire items are chosen for inclusion in a short form based solely on high item-total correlations, the result is often a scale that measures only a narrow portion of the original construct, a phenomenon referred to as the "attenuation paradox" (Loevinger, 1954). Conversely, choosing items that maximize breadth of content may produce scales containing unsatisfactory internal consistency. Therefore, in addition to considering item-total correlations, our decisions regarding inclusion of items were also based on thorough examination of the content of individual items and within-scale factor analysis of the original (standard) scales.

The nature of temperament itself elicited additional concerns. Developmental changes occurring during early childhood create difficulties for temperament measurement. Behaviors indicative of a given trait at an early age are often not informative for measuring the same trait in older children. To address this problem, in creating the short forms, we utilized multiple samples differing in age to ensure that items selected were useful across the intended age range of the questionnaire. In addition, this technique allowed us to avoid one of the more common mistakes of short form developers: basing item inclusion decisions on a single sample, which tends to overestimate the expected reliability of the instrument when used in subsequent studies.

A related consideration concerned missing data. Whereas missing data for particular items is seldom a problem for research on adults, parents often choose the *not applicable* option for certain items when completing the temperament questionnaires on which our short forms are based. For example, when asked whether their child became nervous about going to the dentist, over a third of the parents of 3-year-olds in the sample used to construct the short form indicated that their child had never been observed in that situation. When several items comprise a scale, the issue of missing data is only a minor problem, typically handled by inserting the mean of other item responses or by calculating the scale score as the mean of all completed items. With shorter scales, however, this circumstance is of greater concern, and an initial step in the construction of the short forms was the omission of items with considerable levels of missing data for any age group.

The very short form was constructed in reference to the factor pattern characteristic of the standard form. Factor analysis of the CBQ has consistently resulted in three broad factors (Ahadi et al., 1993; Kochanska, DeVet, Goldman, Murray, & Putnam, 1994; Goldsmith et al., 1997; Rothbart et al., 1994; Rothbart et al., 2001) reminiscent of three of the Big Five (Digman, 1990; Goldberg, 1990) personality dimensions. In U.S. samples, the first factor, Surgency/Extraversion, is characterized by high positive loadings on the Impulsivity, High Intensity Pleasure, and Activity Level scales and strong negative loadings on the Shyness scale. The second factor, Negative Affectivity, is conceptually similar to Neuroticism and is defined by high positive loadings for Sadness, Fear, Anger/Frustration, and Discomfort and negative loadings for Falling Reactivity/Soothability. The third broad factor, Effortful Control, has been compared to Conscientiousness/Constraint and contains high positive loadings for Inhibitory Control, Attentional Control, Low Intensity Pleasure, and Perceptual Sensitivity scales. Positive Anticipation and Smiling and Laughter are inconsistent with respect to primary loadings and often load highly on more than one scale. Although this structure has emerged in exploratory factor analyses of multiple samples, the CBQ was not designed with this structure in mind, and confirmatory factor analyses (CFA) of the scale scores have resulted in inadequate fit to a three-factor model (Rothbart et al., 2001). Because we designed the short form to approximate the specific scales, not the three broad factors, we did not assess the fit of the short form to this model. Because, however, the very short form was created specifically to capture these three broad dimensions, we investigated the fit of the very short form items to the intended structure.

Following the construction of the short and very short forms, we took several steps to assess the psychometric properties of the instruments. In addition to calculating the internal consistency of scores from the short form scales and corrected standard short form correlations, we assessed the reliability of data acquired with the new measures by examining the correspondence between maternal and paternal ratings and assessing longitudinal rank order stability. We also sought to ascertain whether the very short form adequately fit the intended three-factor structure.

In summary, the purpose of this study was to develop short and very short forms of parent-report measures of temperament for children aged 3 to 8 years. We took statistical and theoretical considerations, in addition to issues of comparability across age and patterns of missing data, into account to make item-inclusion decisions. We first describe the samples and procedures used to make decisions regarding item retention. Following this, in Study 1, we present analyses (internal consistency, corrected part–whole correlations, longitudinal rank order stability, cross-informant reliability, and factor structure) conducted using data from a large sample administered the standard form. Study 2 includes analyses of data collected using the short form itself.

SCALE CONSTRUCTION

Samples

We used three samples of children, differing in age, in the construction of the CBQ short and very short forms. The first group was collected by Kochanska et al. (1994) at the University of Iowa and included 171 children (79 girls) with an average age of 39.95 months (*SD* = 11.37; range = 21 to 70). The second and third groups participated in studies conducted by Fagot and Leve (1998) and Fisher (1994) at the Oregon Social Learning Center (OSLC). The second group included 174 children (81 girls) with an average age of 66.65 months (*SD* = 5.55; range = 49 to 92). The third group included 123 children (53 girls) with an average age of 87.67 months (*SD* = 5.58; range = 71 to 101). All groups were predominantly White, with a wide range of socioeconomic status.

Procedure

Construction of the short scales took place in multiple steps. First, we identified the frequency of *not applicable* responses for each item, and we excluded items from consideration for short scales if more than 20% of the respondents in any sample chose the *not applicable* option for the item. We removed three items on this basis. Next, for each scale, we computed Cronbach's alpha and corrected item-total correlations separately for each sample, and we averaged these item-total correlations over the three groups. We then created working scales containing the six items with the highest mean item-total correlations.

A minimum alpha of .65 for data from each scale in each group was desired, as previous work had referred to .65 alphas as satisfactory for a six-item scale (DeVellis, 1991; Francis, Brown, & Philipchalk, 1992). For three scales (Activity Level, Low Intensity Pleasure, and Sadness), we found the scores from the six-item working scales to have $\alpha < .65$ for at least one sample, and we added additional items from the standard scales to these working short scales to increase internal consistency. Scores from the seven-item Activity Level scale met our threshold of .65 for all samples, but the other two seven-item scales still generated αs < .65 for at least one group. Increasing the Low Intensity Pleasure scale to eight items increased alphas beyond our threshold for all groups. Increasing the Sadness scale to eight items did not improve internal consistency, and the seven-item working scale was retained. Thus, the short form contains 12 six-item scales, 2 seven-item scales, and a single eight-item scale.

Next, we performed item-level principal axis factoring on each scale of the standard form. When examination of scree plots indicated a multidimensional scale, we performed oblimin rotation of the factors to identify the items comprising the factors. We then examined the content of items in the working scales with respect to this factor analysis. We then replaced items from the working scales with items not included in the working scales to ensure that all facets of multidimensional scales were represented. For instance, in the two OSLC samples, factor analysis of the item scores for High Intensity Pleasure suggested two factors, one containing items indicating enjoyment of intense, if not risky, activities such as "My child likes rough and rowdy games" and the second containing items indicating thrill-seeking behavior such as "My child likes going down high slides or other adventurous activities." Because the working scale (derived solely on the basis of item-total correlations) contained only two items from the second factor, we replaced an item loading highly on the first factor with an item loading primarily on the second factor. We then computed alpha coefficients for scores corresponding to the revised working scales. In two cases, it was not possible to represent all facets of the multidimensional scale while maintaining acceptable internal consistency (i.e., $\alpha > .65$ for all groups). For these two scales, the first factors to emerge were deemed most representative of the standard form scale, and we used items loading primarily on these factors in the short form scales. For the Discomfort scale, the first factor to emerge in all three groups referred to reactions to pain (e.g., cuts and bruises, being cold or wet, being ill with a cold) and the other indexed reactions to intense stimuli (e.g., bright lights, loud sounds, rough materials). The short version of this scale includes only items concerning pain reactions. The Attentional Control scale contained a factor corresponding to ability to maintain attentional focus and a second referring to facility in willfully shifting attention. The short version includes only the former.

The final step in scale construction involved a thorough content analysis of the items in the revised working scales. Our goal was to ensure breadth of item content while maintaining adequate internal consistency. When more than one item in a working scale referred to children's behavior in the same (or a highly similar) situation, we removed one of the items and replaced it with an item that did not share content with any items in the working scale. We undertook this step only if it did not result in a short scale with $\alpha < .65$ for any sample.

We developed the very short form of the CBQ for researchers interested in efficiently obtaining scores for only the three factors. The goal was to create three orthogonal scales reflecting the broad content of the factors. To select items, we created scores for each of the three factors by averaging standard scale scores corresponding to the factor (e.g., an Effortful Control score was created by averaging scale scores for Attention Control, Inhibitory Control, Perceptual Sensitivity, and Low Intensity Pleasure). We then examined items that had been selected for inclusion in the short form in relation to these scores. We considered items exhibiting large correlations with their associated factor, and small correlations with the other two factors, for the very short form. We retained two or three items from each scale for the very short form (e.g., the very short Negative Affect scale contains two Frustration items, three Discomfort items, two Soothability items, three Sadness items, and two Fear items).

STUDY 1: SHORT AND VERY SHORT FORMS EXTRACTED FROM STANDARD FORM

Samples

We acquired data sets by contacting, through email or postal mail, researchers who had requested information regarding the CBQ between 1997 and 2000, and we obtained data from the following five North American sources: from Stephanie Carlson at University of Washington, 245 (129 female) children with an average age of 49.00 months (*SD* = 6.47; range = 38 to 66); from Lucy LeMare at Simon Fraser University, 129 (49 female) children with an average age of 70.60 months (*SD* = 6.58; range = 60 to 83); from Grazyna Kochanska at University of Iowa, 99 (48 female) children with an average age of 45.28 months (*SD* = .72; range = 44 to 50); from Megan Gunnar at University of Minnesota, 60 (31 female) children with an average age of 74.38 months (*SD* = 1.33; range = 71 to 78); and from Mary Rothbart at University of Oregon, 57 (28 female) children, all 36 months of age. The entire sample included 590 (285 female) children with an average age of 54.42 months (*SD* = 13.57; range = 36 to 83). All samples were primarily White and of middle to upper socioeconomic status.

The children in Kochanska's sample had previously been tested at an average age of 32.80 months (*SD* = .53; range = 32 to 34), with CBQs completed by both mothers and fathers. We used only the mother reports from the second collection in our internal consistency, convergent validity, and factor structure analyses. We used the mother report data from the earlier collection and the father data from both time points to assess rank order stability and cross-informant agreement for the standard and short forms. All mothers who completed the CBQ for the second collection did so at the earlier time point. Forms were completed by 94 fathers during the first assessment. Of these fathers, 82 completed CBQs at the second visit. In addition, 2 fathers completed the measure for the second collection only.

Results

Internal consistency. Alpha coefficients obtained for the scales of the standard and short forms are shown in Table 1. Standard errors for these alphas, which we calculated using a method described by Iacobucci and Duhachek (2003), were all less than .01. Alpha coefficients for the short form scales were approximately .06 lower, on average, than the corresponding values for standard scales. Of the 15 short scales, 11 achieved alphas over .70, and the alpha for only 1 scale, Sadness, was below .65. Whereas alphas for 13 scales decreased from the standard to short forms, internal consistency of the Attention Focusing and Discomfort short scales was greater than the corresponding standard scales. Alphas for the Surgency, Negative Affect, and Effortful Control scales of the very short form equaled .75, .72, and .74, respectively.

Standard to short form relations. To assess the correspondence between the standard and short scales, we applied Levy's (1967) correction. 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 (Petrides, Jackson, Furnham, & Levine, 2003).[1] As shown in Table 1, corrected correlation coefficients were above .70 for 12 of the 15 scales, with only 1 scale, Sadness, attaining a correlation below .65. We also created standard form scores for Surgency, Negative Affect, and Effortful Control by summing scores of all items from scales associated with each of the three factors. Corrected standard to very short correlations utilizing these scales were .83, .75, and .83 for Surgency, Negative Affect, and Effortful Control, respectively.

Longitudinal stability. Rank order stability correlations for the standard and short forms from approximately 33 to 45 months can be found in Table 1. Stability coefficients for the short form scales were approximately .05 lower, on average, than the corresponding correlations for standard scales for mother report and approximately .04 lower for father report.

Using the maternal ratings, stability correlations for scores from the very short form scales were .73, .70, and .63

[1]A computer program for applying the Levy (1967) correction can be obtained from P. Barrett's Web site at http://www.pbarrett.net/shortform.htm.

TABLE 1
Internal Consistency, Interrater Reliability, Longitudinal Stability of, and Correlations Between CBQ Standard and Short-Form Scales in Study 1

			Interrater Reliability		*Rank Order Stability*[a]		
Scale	*No. Items*	α	*33 Months*	*45 Months*	*Mother*	*Father*	*Short to Standard Corrected r*
Activity Level							.79
Standard	13	.81	.40***	.40***	.77***	.73***	
Short	7	.75	.38***	.45***	.80***	.64***	
Anger/Frustration							.75
Standard	13	.81	.39***	.47***	.73***	.59***	
Short	6	.76	.40***	.51***	.70***	.52***	
Approach/Positive Anticipation							.71
Standard	13	.74	.39***	.17	.70***	.54***	
Short	6	.65	.35***	.13	.54***	.56***	
Attentional Focusing							.70
Standard	9	.73	.48***	.53***	.65***	.71***	
Short	6	.75	.47***	.53***	.61***	.71***	
Discomfort							.72
Standard	12	.72	.42***	.46***	.72***	.70***	
Short	6	.79	.46***	.59***	.74***	.59***	
Soothability							.72
Standard	13	.75	.49***	.47***	.59***	.54***	
Short	6	.73	.43***	.41***	.53***	.56***	
Fear							.69
Standard	12	.69	.52***	.49***	.57***	.55***	
Short	6	.68	.45***	.55***	.58***	.56***	
High Intensity Pleasure							.75
Standard	13	.79	.37***	.49***	.76***	.60***	
Short	6	.72	.39***	.40***	.71***	.60***	
Impulsivity							.77
Standard	13	.81	.48***	.43***	.75***	.55***	
Short	6	.72	.40***	.42***	.75***	.51***	
Inhibitory Control							.79
Standard	13	.83	.58***	.62***	.78***	.72***	
Short	6	.72	.47***	.49***	.70***	.64***	
Low Intensity Pleasure							.66
Standard	13	.72	.36***	.55***	.78***	.60***	
Short	8	.69	.33***	.50***	.74***	.41***	
Perceptual Sensitivity							.73
Standard	12	.77	.13	.17	.58***	.50***	
Short	6	.73	.08	.26**	.55***	.49***	
Sadness							.62
Standard	12	.69	.46***	.33***	.71***	.54***	
Short	7	.61	.26**	.34***	.65***	.40***	
Shyness							.88
Standard	13	.93	.55***	.53***	.74***	.75***	
Short	6	.85	.51***	.43***	.63***	.74***	
Smiling and Laughter							.77
Standard	13	.79	.27***	.34***	.71***	.59**	
Short	6	.71	.18*	.31***	.56***	.62***	

Note. 33-month interrater reliability $n = 98$; 46-month interrater reliability $n = 84$; mother report stability $n = 100$; father report stability $n = 82$. Standard error < .01 for all alphas. CBQ = Children's Behavior Questionnaire.
[a]33 to 46 months.
*$p < .10$. **$p < .05$. ***$p < .01$.

for Surgency, Negative Affect, and Effortful Control, respectively. Corresponding coefficients for the paternal ratings were .62, .61, and .64, respectively.

Cross-informant reliability. Pearson's correlations regarding agreement between mother and father reports for the standard and short forms are shown in Table 1. At 33 and 45 months, respectively, parental agreement correlations for short form scales were approximately .05 and .01 lower, on average, than for standard scales. Interparent agreement for both standard and short forms was particularly low for Perceptual Sensitivity at both ages and for Approach/Positive Anticipation at 45 months.

For the very short form at 33 months, correlations between mother and father ratings equaled .45, .36, and .22 for Surgency, Negative Affect, and Effortful Control, respectively. At 45 months, the respective cross-informant correlations were .57, .52, and .33, respectively.

Very short form factor structure. We performed maximum likelihood CFA on the covariance matrix of the 36 raw item scores to assess whether the very short form data would fit the intended orthogonal three-factor model. We allowed each item to load on their specified scale factor and constrained to have zero loadings with the other two factors. We did not allow the latent factors to correlate with one another. Following Kline (1998), we utilized multiple fit indexes. In an initial model, we did not allow item error terms to correlate. The comparative fit index (CFI) and Tucker–Lewis index (TLI) were .96, indicating acceptable fit (Hu & Bentler, 1999). With $\chi^2(594, N = 590) = 2{,}599$, the ratio of χ^2 to $df = 4.38$, and the root mean square error of approximation (RMSEA) was .076 (90% confidence interval [CI] = .073 to .079), both indicating that the fit could be improved (Hu & Bentler, 1999; Kline, 1998). In a second model, we allowed error terms for items taken from the same standard scale (e.g., the three Activity Level items) to correlate. All indexes suggested at least marginally adequate fit to this model: CFI = .99, TLI = .98, $\chi^2(562, N = 590) = 1{,}370$ (χ^2:$df = 2.44$), RMSEA = .049 (CI = .046 to .053). Item loadings are shown in Table 2.

The CFA models assumed orthogonal factors. We calculated Pearson's correlations to assess the degree of orthogonality of the scales themselves. We obtained small but statistically significant correlations between Surgency and Effortful Control, $r(590) = -.19$, $p < .01$, and Surgency and Negative Affect, $r(590) = -.08$, $p < .05$. Negative Affect and Effortful Control were not significantly correlated, $r(590) = -.02$, $p > .10$.

TABLE 2
Standardized Parameter Estimates of CBQ VSF Items in Study 1

VSF Item No.	*CBQ Scale*	*Negative Affect*	*Surgency*	*Effortful Control*
2	Anger	.54		
32	Anger	.53		
29	Discomfort	.43		
8	Sadness	.42		
17	Sadness	.41		
5	Discomfort	.40		
14	Soothability	.39		
23	Soothability	.38		
11	Fear	.31		
20	Discomfort	.25		
26	Fear	.24		
35	Sadness	.23		
7	Impulsivity		.97	
19	Impulsivity		.95	
31	Impulsivity		.62	
34	Shyness		.43	
10	Shyness		.43	
22	Shyness		.32	
1	Activity Level		.27	
13	Activity Level		.25	
4	High Intensity Pleasure		.24	
25	Activity Level		.21	
28	High Intensity Pleasure		.17	
16	High Intensity Pleasure		.11	
18	Inhibitory Control			.65
30	Inhibitory Control			.58
6	Inhibitory Control			.57
3	Attention Focusing			.40
21	Low Intensity Pleasure			.40
12	Perceptual Sensitivity			.36
24	Perceptual Sensitivity			.36
15	Attention Focusing			.32
36	Perceptual Sensitivity			.32
9	Low Intensity Pleasure			.30
27	Attention Focusing			.28
33	Low Intensity Pleasure			.26

Note. CBQ = Children's Behavior Questionnaire; VSF = Very Short Form.

Discussion

Although short versions of measurement scales inevitably lose breadth in content and/or exhibit lower internal consistency than their parent scales, the results of Study 1 suggest that our short form of the CBQ is an acceptable measure for researchers wishing to measure a large variety of attributes while minimizing participant time. A total of 11 scales achieved alphas greater than .70 in the Study 1 data, and alpha for only 1 scale, Sadness, was below .65. Although .70 is widely considered a benchmark for good internal consistency, DeVellis (1991) rated alphas of .60 as undesirable, but not unacceptable. Although instruments devised for clinical purposes typically strive for greater levels of internal consistency, the consistency characterizing the short form scales appears adequate for research purposes, as they are higher than several reported for longer scales of previously developed temperament questionnaires (Fullard, McDevitt, & Carey, 1984; McDevitt & Carey, 1978). We also note that internal consistency was not our sole concern during scale refinement. Although high item-total correlations were used to develop the initial working scales of the short form, several items were removed and replaced to create broader, and thus more valid, scales (Loevinger, 1954).

Because a critical aspect of temperament is longitudinal stability, it was important to demonstrate that rank order stability coefficients for scores on the short and very short forms were comparable to those obtained with the standard scales. Patterns of stability were consistent between the standard and short forms, with scales exhibiting greater or lesser levels of stability in the standard form continuing to do so in the short form. These results were found using data from both mothers and fathers. Although these stability coefficients were considerably smaller than test–retest reliability estimates typically sought by questionnaire creators, they were higher, on average, than stability correlations for temperament over comparable intervals during early childhood that has been reported in previous studies (Earls & Jung, 1987; Guerin & Gottfried, 1994).

For the majority of scales, the level of cross-rater reliability at both 33 and 46 months for the standard, short, and very short forms was consistent with interparent agreement for child temperament in other studies (see review by Slabach, Morrow, & Wachs, 1991). The Perceptual Sensitivity and Approach/Positive Anticipation scales (at 46 months only), however, did not exhibit expected levels of cross-rater reliability. This problem is not specific to the short form, as the standard scales also demonstrated questionable interparent agreement, but may suggest either subjectivity in parental ratings of child behavior indicating these traits, or alternatively, tendencies for fathers and mothers to elicit different behavior from the same child (see Bates, 1989; Rothbart & Goldsmith, 1985). Regardless, researchers are cautioned regarding their interpretation of scores obtained with these scales.

The three broad scales of the very short form exhibited good internal consistency and high corrected standard–short correlations. Furthermore, although prior scale-level CFA of the standard CBQ achieved fit to a three-factor model only after substantial use of modification indexes (Rothbart et al., 2001), item-level analyses of the very short form in Study 1 achieved acceptable fit with less drastic alterations.

STUDY 2: SHORT FORM ADMINISTERED

In Study 2, we address two problems inherent in the data utilized in the questionnaire creation and in Study 1. First, the data used in the construction of the short forms and in Study 1 were collected using the standard CBQ, leaving open the possibility that the psychometric properties of the short and very short forms might be lowered when these instruments are administered instead of extracted from the standard form. A second limitation concerns the nature of the samples. The majority of the data used to develop and assess the forms were drawn from the Midwest and Northwest United States, predominantly from rural and suburban areas. In Study 2, we instead used data from more ethnically and financially diverse samples to which we administered the short, not the standard, form of the CBQ.

Sample

Three data sets were acquired. Sample 1 was acquired from James Victor at Hampton University and consisted of 138 children (78 female) between the ages of 3 and 8 (average age = 68.00 months, *SD* = 20.67). In contrast to the primarily White composition of the Study 1 samples, this sample was 24% White, 69% African American, and 7% racially mixed. Sample 2, contributed by Elizabeth Anson, Robert Cole, Harriet Kitzman, and Kimberly Sidora-Arcoleo at the University of Rochester Medical Center, consisted of 289 (136 female) 3-year-olds, 34% of whom were White, 48% were African American, and 18% were non-African American or racially mixed. Of this sample, 49% were identified as living in poverty based on an analysis of income to needs. Sample 3 was acquired from Stephanie Carlson at the University of Washington and contained 169 children (78 female), predominantly White, ranging in age from 39 to 60 months (average age = 50.17 months, *SD* = 4.83). Because we wished to assess the impact of sample characteristics, we present internal consistency analyses separately for the three samples.

Results

Internal consistency. Alpha coefficients obtained for the short and (extracted) very short scales are shown in Table 3. Internal consistency estimates of the short form scales in the Sample 3 data were very similar to those obtained from extracted short form data in Study 1, with 11 of 15 scales that exhibited alphas over .70 and 14 alphas over .60. The exception was the Sadness scale, which had the lowest alpha in Study 1 and achieved unacceptably low alpha of .46 in the Sample 3 data set. This scale also exhibited poor internal consistency ($\alpha = .43$) in Sample 2.

In contrast to the acceptable internal consistency in Study 1, and Sample 3 of Study 2, both of which were composed primarily of mid-income to upper income White respondents, several short form scales exhibited undesirable or unacceptable internal consistency when administered to more diverse samples. As shown in Table 3, seven alphas in the Sample 1 data were below .70, and three were below .60. In the Sample 2 data, 13 alphas were below .70, and 3 were be-

TABLE 3
Internal Consistency (Alphas) of Short Form and Very Short Form Scales in Three Study 2 Samples

	Sample 1[a]	*Sample 2*[b]	*Sample 3*[c]
Short Form Scale			
Activity Level	.74	.65	.72
Anger/Frustration	.78	.72	.69
Approach/Positive Anticipation	.58	.57	.70
Attentional Focusing	.73	.70	.70
Discomfort	.70	.69	.82
Soothability	.71	.67	.80
Fear	.54	.64	.60
High Intensity Pleasure	.70	.66	.76
Impulsivity	.54	.62	.74
Inhibitory Control	.62	.68	.72
Low Intensity Pleasure	.82	.60	.68
Perceptual Sensitivity	.69	.60	.73
Sadness	.61	.43	.46
Shyness	.79	.82	.87
Smiling and Laughter	.64	.55	.77
Very Short Form Scale			
Surgency	.73	.70	.76
Negative Affect	.66	.70	.67
Effortful Control	.78	.62	.77

Note. Standard errors < .025.
[a]N = 138. [b]N = 289. [c]N = 169.

low .60. Curiously, this pattern was not evident for the scales of the very short form.

To explore the possibility that the compromised internal consistency of the short form scales in these data sets was related to sample characteristics, we divided the samples by racial status and ran analyses separately for Whites and African Americans (racially mixed subsamples were excessively small and too heterogeneous to permit separate analyses). Because we had income information for Sample 2, we additionally performed analyses on groups split by poverty status. The results of these analyses, shown in Table 4, support the contention that the short form scales perform less optimally in impoverished and African American samples. Whereas in the White groups, 13 of 30 alphas calculated were lower than .70, and 2 were below .60, the corresponding numbers for the African American groups were 24 and 13, respectively. Socioeconomic status affected internal consistency in a similar manner: For the Sample 2 subsample above the poverty line, 8 alphas were below .70 and 2 were below .60, and the corresponding numbers for the impoverished subsample were 12 and 5, respectively.

Very short form factor structure. To maximize power, we combined the three Study 2 samples for CFA of the very short form. Similar to the Study 1 analyses, when we did not allow the item errors to correlate, CFI = .96, TLI = .96, $\chi^2(594, N = 594) = 2,548$ ($\chi^2{:}df = 4.29$), RMSEA = .074 (CI = .072 to .077). When we allowed error terms for items from the same scale to correlate, all indexes suggested at least a marginally acceptable fit: CFI = .98, TLI = .98, $\chi^2(561, N = 594) = 1,491$ $\chi^2{:}df = 2.66$, RMSEA = .053 (CI = .050 to .056).

Small correlations between the scale scores were evident. As in Study 1, Surgency and Effortful Control were negatively correlated, $r(594) = -.10$, $p < .05$. Negative Affect and Effortful Control were also negatively correlated, $r(594) = -.10$, $p < .05$. Negative Affect and Surgency were not significantly correlated, $r(594) = .01$, $p > .10$.

Discussion

When administered to samples similar to those used to construct and initially assess the short form, internal consistency estimates of the short form scales closely approximated those generated when data were extracted from the standard form of the CBQ. In addition, regardless of sample characteristics, alphas for the very short form scales (extracted from the short form) were similar to those obtained when extracted from the standard form, and the fit of CFA models to very short form data was adequate. Internal consistency estimates of the scales were considerably lower when analyses were restricted to African American and low-income samples, although the majority of scales continued to demonstrate alphas higher than .60, considered by DeVellis (1991) to be the threshold for acceptable internal consistency.

GENERAL DISCUSSION

The purpose of this investigation was to develop and evaluate the psychometric characteristics of short and very short forms of the CBQ. In addition to considering internal consistency of scores while selecting items for these forms, we made efforts to include items that were completed by the

TABLE 4
Alphas and Standard Errors for Study 2 Samples Divided by Race and Poverty Status

	Sample 1				*Sample 2*							
	White[a]		*African American*[b]		*White*[c]		*African American*[d]		*Not Poverty*[e]		*Poverty*[f]	
Short Form Scale	α	*SE*	α	*SE*	α	*SE*	α	*SE*	α	*SE*	α	*SE*
Activity Level	.86	.02	.68	.01	.72	.01	.60	.01	.70	.01	.62	.02
Anger/Frustration	.82	.02	.77	.01	.79	.01	.67	.02	.72	.01	.72	.01
Approach/Positive Anticipation	.61	.04	.56	.02	.66	.01	.55	.01	.65	.02	.53	.02
Attentional Focusing	.66	.03	.74	.01	.79	.01	.64	.02	.73	.01	.70	.02
Discomfort	.85	.02	.63	.02	.72	.02	.64	.02	.76	.01	.64	.02
Soothability	.70	.03	.71	.02	.73	.02	.63	.02	.65	.01	.68	.02
Fear	.46	.06	.56	.03	.77	.02	.48	.02	.65	.02	.64	.01
High Intensity Pleasure	.78	.02	.67	.02	.70	.02	.66	.02	.70	.02	.65	.02
Impulsivity	.77	.03	.41	.03	.68	.02	.50	.02	.68	.02	.48	.03
Inhibitory Control	.74	.03	.57	.02	.71	.02	.64	.02	.70	.01	.65	.02
Low Intensity Pleasure	.68	.02	.84	.01	.62	.02	.60	.01	.61	.01	.63	.02
Perceptual Sensitivity	.69	.03	.69	.02	.61	.02	.56	.02	.59	.02	.58	.02
Sadness	.68	.03	.60	.02	.59	.02	.35	.03	.42	.03	.49	.03
Shyness	.87	.02	.74	.02	.86	.01	.80	.01	.81	.01	.83	.01
Smiling and Laughter	.61	.04	.63	.02	.66	.02	.53	.02	.64	.02	.55	.02

[a]*n* = 32. [b]*n* = 94. [c]*n* = 96. [d]*n* = 136. [e]*n* = 140. [f]*n* = 133.

large majority of participants and that approximated the domain of items included within the parent scale. Rather than relying on a single sample during the item selection process, we used three samples differing in age. In subsequent analysis, when we used data extracted from the standard CBQ and collected with the short form, the large majority of short scales showed adequate levels of internal consistency and were highly correlated with the standard form scales. In addition, the short scales were nearly as consistent across time and raters as were the standard scales. The three scales of the very short form exhibited acceptable internal consistency across all samples, and CFA demonstrated marginally acceptable fit to the intended three-factor structure.

These new measures give researchers a great deal of flexibility in choosing an instrument that corresponds to their specific needs. The very short form may be most useful for investigators whose primary research interests lie in other areas but also aspire to efficiently assess established broad dimensions of temperament to control for temperament in their analyses or to address secondary temperament-related questions. The 195-item standard measure takes approximately 1 hr to complete. Because the very short form contains less than a fifth of the items from the standard CBQ, it is estimated that participants will be able to complete the entire instrument in less than 15 min. As such, the very short CBQ can be easily included in a larger battery of measures.

The short form is appropriate for use by investigators who lack the time to administer the standard CBQ but still desire assessment of a wide variety of traits to examine more specific questions concerning temperament. Because the short form is less than half the length of the standard instrument, approximately 30 min of participants' time will be saved. Although the short form will save research participants time and energy, abbreviated measures inevitably result in some loss of important information, and it is recommended that researchers use the longer form when it is feasible. Researchers should also consider the alternative strategy of administering only a selected subset of CBQ scales when their research question is strongly focused. This alternative maintains the valuable characteristics of the standard scales while reducing time and energy demands on participants but may leave out important information on temperament in the scales that have been discarded.

Developmental researchers interested in specific dimensions of negative affect and effortful control are particularly encouraged to use the longer forms of the Discomfort, Sadness, and Attentional Focusing scales if these emotions are central to their query. The short Discomfort and Attention Focusing scales actually exhibited higher internal consistency than their parent scales, but this is likely due to a narrowing of item content, which constrains the content validity of the short scales. The short Sadness scale is similar in content to the longer scale. The parent scale itself, however, suffered from low internal consistency, and reducing the number of items has reduced the reliability to unsatisfactory levels for some groups. Despite the problems with these scales, they continued to exhibit cross-age and cross-informant consistency at levels similar to the other scales. Investigators should also be careful in interpreting scores for Perceptual Sensitivity and Approach/Positive Anticipation due to low levels of interparent agreement for these scales.

As noted by Thompson (1994) and others (Gronlund & Lind, 1990), reliability is not a characteristic of a scale itself but is a property of data and may fluctuate in relation to sample characteristics. The results of Study 2 suggest that caution is warranted regarding the use of the short (but not very short) form in samples that are predominantly composed of African American or low-income participants because internal consistency was substantially lower for several scales when analyses were restricted to samples of these nature. These results were not anticipated, but generated only in an effort to discern the cause of low internal consistency in the samples acquired. Multiple factors may contribute to these results. One potential explanation concerns differences in dialect. Gopaul-McNicol, Reid, and Wisdom (1998) reviewed literature that has indicated the limitations of a number of psychoeducational instruments for individuals who speak subcultural dialects such as Ebonics, and it is plausible that dialectal rules led to alternative interpretations of questionnaire items, lowering internal consistency of scores from minority and low-income participants. Educational differences between groups may have similarly led to differing item interpretations. Because participants were not read the items but completed the instruments on their own, reading difficulties may have contributed to inaccurate responding for some items. Interestingly, in both samples for which data could be analyzed separately by minority and poverty status, scales measuring Activity Level and Impulsivity (and to a lesser extent, Discomfort) exhibited the strongest decreases in internal consistency. It may be the case that these traits may be perceived as more context-specific in low-income or minority samples. For instance, space limitations in low-income housing may create a condition wherein a child who is highly active when outside may not move about actively indoors, leading to low correspondence between items assessing Activity Level during inside and outside play. It should be noted that although the reliability of some scales was compromised among these subsamples, the majority of short form scales continued to evince adequate internal consistency.

In constructing and testing the short and very short forms of the CBQ, we have attempted to avoid several "sins" associated with short-form development (Smith, McCarthy, & Anderson, 2000). We remain guilty, however, of one commonly committed sin. As noted in our discussion of Study 1, due to shared error, estimates of agreement between short and standard forms are exaggerated when short form data are extracted from data collected with a standard form. Although we statistically controlled for shared error, future studies administering both forms to the same sample would more accurately assess agreement between the short and standard forms. A similar critique can be made of our analysis of the very short form: Because this form has not yet been administered but only extracted from standard and short form data, there is a possibility that the psychometric properties of the very short form might be lowered when these instruments are administered. Conversely, however, it could be argued that respondents may be more conscientious when filling out the very short form, resulting in better statistical properties. In addition, future investigations involving relations between the CBQ dimensions and other psychological phenomena, using the short and very short forms themselves, are necessary to further assess the validity of these instruments.

ACKNOWLEDGMENTS

This research was supported by funds allocated to S. P. Putnam from National Institute of Mental Health (NIMH) Grant 5 T32 MH1893, awarded to the University of Oregon, and by NIMH Grants MH01471 and MH43361 awarded to M. K. Rothbart. Preliminary findings were presented at the biennial meeting of the Society for Research in Child Development in Tampa, Florida, April 2003. Copies of the standard, short, and very short forms of the CBQ may be acquired from S. P. Putnam's Web site at http://www.bowdoin.edu/~sputnam/rothbart-temperament-questionnaires/. We express our gratitude to the researchers who generously donated their CBQ data to this effort, particularly Stephanie Carlson and Grazyna Kochanska who contributed multiple data sets.

REFERENCES

Ahadi, S. A., Rothbart, M. K., & Ye, R. M. (1993). Children's temperament in the U.S. and China: Similarities and differences. *European Journal of Personality, 7,* 359–377.

Aksan, N., Goldsmith, H. H, Smider, N. A., Essex, M. J., Clark, R., Hyde, J. S., et al. (1999). Derivation and prediction of temperamental types among preschoolers. *Developmental Psychology, 35,* 958–971.

Bates, J. E. (1989). Concepts and measure of temperament. In G. A. Kohnstamm, J. E. Bates, & M. K. Rothbart (Eds.), *Temperament in childhood* (pp. 3–16). Chichester, England: Wiley.

Derryberry, D., & Rothbart, M. K. (1988). Arousal, affective, and attentional components of adult temperament. *Journal of Personality and Social Psychology, 55,* 953–966.

DeVellis, R. R. (1991). *Scale development: Theory and applications.* Thousand Oaks, CA: Sage.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Donzella, B., Gunnar, M. R., Krueger, W. K., & Alwin, J. (2000). Cortisol and vagal tone responses to competitive challenge in preschoolers: Associations with temperament. *Developmental Psychobiology, 37,* 209–220.

Earls, F., & Jung, K. G. (1987). Temperament and home environment characteristics as causal factors in the early development of childhood psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry, 26,* 491–498.

Eisenberg, N., Fabes, R. A., Guthrie, I. K., & Murphy, B. C. (1996). The relations of regulation and emotionality to problem behavior in elementary school children. *Development and Psychopathology, 8,* 141–162.

Eisenberg, N., Fabes, R. A., Karbon, M., & Murphy, B. C. (1996). The relations of children's dispositional prosocial behavior to emotionality, regulation, and social functioning. *Child Development, 67,* 974–992.

Eisenberg, N., Fabes, R. A., Shepard, S. A., Guthrie, I. K., Murphy, B. C., & Reiser, M. (1999). Parental reactions to children's negative emotions: Longitudinal relations to quality of children's social functioning. *Child Development, 70,* 513–534.

Fabes, R. A., Eisenberg, N., Jones, S., Smith, M., Guthrie, I., Poulin, R., et al. (1999). Regulation, emotionality, and preschoolers' socially competent peer interactions. *Child Development, 70,* 432–442.

Fagot, B. I., & Leve, L. D. (1998). Teacher ratings of externalizing behavior at school entry for boys and girls: Similar early predictors and different correlates. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 39,* 555–566.

Fisher, P. A. (1994). Temperament goodness of fit and psychosocial adjustment in children. *Dissertation Abstracts International, 54*(11B), 5941.

Francis, L. J., Brown, L. B., & Philipchalk, R. (1992). The development of an abbreviated form of the Revised Eysenck Personality Questionnaire (EPQR–A): Its use among students in England, Canada, the USA, and Australia. *Personality and Individual Differences, 13,* 443–449.

Fullard, W., McDevitt, S. C., & Carey, W. B. (1984). Assessing temperament in one- to three-year-old children. *Journal of Pediatric Psychology, 9,* 205–217.

Geurin, D. W., & Gottfried, A. W. (1994). Developmental stability and change in parent reports of temperament: A ten-year longitudinal investigation from infancy through preadolescence. *Merrill-Palmer Quarterly, 40,* 334–355.

Gopaul-McNicol, S., Reid, G., & Wisdom, C. (1998). The psychoeducational assessment of ebonics speakers: Issues and challenges. *Journal of Negro Education, 67,* 16–24.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Goldsmith, H. H., Buss, K. A., & Lemery, K. S. (1997). Toddler and childhood temperament: Expanded content, stronger genetic evidence, new evidence for the importance of environment. *Developmental Psychology, 33,* 891–905.

Gruonlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Guthrie, I. K., Eisenberg, N., Fabes, R. A., Murphy, B. C., Holmgren, R., Mazsk, P., et al. (1997). The relations of regulation and emotionality to children's situational empathy-related responding. *Motivation and Emotion, 21,* 87–108.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology, 13,* 478–487.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Kochanska, G. (1997). Multiple pathways to conscience for children with different temperaments: From toddlerhood to age 5. *Developmental Psychology, 33,* 228–240.

Kochanska, G., DeVet, K., Goldman, M., Murray, K., & Putnam, S. P. (1994). Maternal reports of conscience development and temperament in young children. *Child Development, 65,* 852–868.

Lengua, L. J., West, S. G., & Sandler, I. N. (1998). Temperament as a predictor of symptomatology in children: Addressing contamination of measures. *Child Development, 69,* 164–181.

Lengua, L. J., Wolchik, A. W., Sandler, I. N., & West, S. G. (2000). The additive and interactive effects of parenting and temperament in predicting adjustment problems of children of divorce. *Journal of Clinical Child Psychology, 29,* 232–244.

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

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

McDevitt, S. C., & Carey, W. B. (1978). The measurement of temperament in 3–7 year old children. *Journal of Child Psychology and Psychiatry, 19,* 245–253.

Mills, R. S. L. (1998). Paradoxical relations between perceived power and maternal control. *Merrill-Palmer Quarterly, 44,* 523–537.

Murphy, B. C., Eisenberg, N., Fabes, R. A., Shepard, S., & Guthrie, I. K. (1999). Consistency and change in children's emotionality and regulation: A longitudinal study. *Merrill-Palmer Quarterly, 45,* 413–444.

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.

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

Rothbart, M. K., Ahadi, S. A., & Hershey, K. L. (1994). Temperament and social behavior in childhood. *Merrill-Palmer Quarterly, 40,* 21–39.

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.

Rothbart, M. K., & Derryberry, D. (1981). Development of individual differences in temperament. In M. E. Lamb & A. L. Brown (Eds.), *Advances in developmental psychology* (Vol. 1, pp. 37–86). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Rothbart, M. K., & Goldsmith, H. H. (1985). Three approaches to the study of infant temperament. *Developmental Review, 5,* 237–260.

Schaughency, E., & Fagot, B. I. (1993). The prediction of adjustment at age 7 from activity level at age 5. *Journal of Abnormal Child Psychology, 21,* 29–50.

Schwebel, D. C., & Plumert, J. M. (1999). Longitudinal and concurrent relations among temperament, ability estimation, and injury proneness. *Child Development, 70,* 700–712.

Silverman, I. W., & Ippolito, M. F. (1995). Maternal antecedents of delay ability in young children. *Journal of Applied Developmental Psychology, 16,* 569–591.

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: Plenum.

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

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

Tomlinson, P. S., Harbaugh, B. L., & Anderson, K. H. (1996). Children's temperament at 3 months and 6 years old: Stability, reliability, and measurement issues. *Issues in Comprehensive Pediatric Nursing, 19,* 33–47.

Samuel P. Putnam
Department of Psychology
Bowdoin College
6900 College Station
Brunswick, ME 04011–8469
Email: sputnam@bowdoin.edu

Received October 10, 2005
Revised February 17, 2006