Stability of Laboratory-Assessed Temperamental Emotionality Traits
From Ages 3 to 7

C. Emily Durbin
Northwestern University

Elizabeth P. Hayden
University of Western Ontario

Daniel N. Klein and Thomas M. Olino
Stony Brook University

A key component of temperament models is the presumed temporal stability of temperament traits. Although a substantial literature using parent report measures has addressed this claim, very few investigations have examined the stability of temperament using alternative measurement strategies, particularly those that involve direct assessment of emotional expressions. This study reports on the relative stability and heterotypic continuity of temperament traits measured via laboratory tasks and maternal report in a sample of children assessed at ages 3, 5, and 7, focusing on Positive Emotionality and Negative Emotionality. Relative stability of Positive Emotionality and Negative Emotionality traits ranged from moderate to high for laboratory and maternal report measures. Measures of emotional expressions exhibited levels of stability comparable to or higher than traits defined by other behavioral patterns (e.g., sociability and engagement).

Keywords: emotion, temperament, stability

Temperament has been defined as individual differences in emotional reactivity and regulation (Goldsmith, Lemery, Askan, & Buss, 2000; Rothbart & Derryberry, 1981) that organize behavior and influence responsivity to the environment (Caspi, 1998). A key component of modern models of temperament is the claim that these traits exhibit continuity across time (Goldsmith et al., 1987). Thus, although the surface behaviors emerging from a latent temperament trait may take developmentally distinct forms, it is presumed that individual differences in underlying levels of the latent traits are relatively stable, producing continuity between traits observed in childhood and adulthood and stability across the life span (Caspi & Roberts, 2005).

The stability of adult personality traits, including those defined by emotional reactivity, is well established. Roberts and DelVecchio (2000) concluded from a meta-analysis of longitudinal studies of personality traits that the average trait consistency was .55 across a 5-year period and .52 over a 10-year period. Although stability indices tend to increase with age among adult samples (Roberts & DelVecchio, 2000), little is known about the effects of age on trait consistency in childhood and adolescence, with the exception that stability tends to be lower before age 3 (Caspi, 2000; Pedlow, Sanson, Prior, & Oberklaid, 1993).

Evidence is accumulating that childhood temperament traits exhibit continuity with personality measured in adulthood and with conceptually related markers of psychosocial adjustment (Caspi, 2000; Caspi et al., 2003; Shiner, 2000). Specifically, there is a growing consensus that the superfactors of Positive Emotionality (PE)/Extraversion, and Negative Emotionality (NE)/Neuroticism represent core traits present across the life span (John, Caspi, Robins, Moffitt, & Stouthamer-Loeber, 1994; Rothbart & Bates, 1998; Tellegen, 1985). In childhood, these superfactors appear to be marked by the following lower order facets: (a) PE: positive affect, sociability, social inhibition, and dominance and (b) NE: fear, sadness, anger, and discomfort (Rothbart, Ahadi, & Evans, 2000; Shiner, 1998). Activity level/energy has been linked to both PE and NE in child temperament models (Rothbart & Bates, 1998), but has generally been identified as a marker of PE in the adult literature (Watson & Clark, 1997). In addition, both child and adult models propose that PE and NE are independent of one another (Clark & Watson, 1999; Rothbart, Ahadi, & Evans, 2000). A number of theorists have argued that PE and NE form the “core” of personality/temperament (Rothbart, Ahadi, et al., 2000; Tellegen, 1985; Watson & Tellegen, 1985) and that these superfactors are both earlier appearing and more stable into adulthood than narrow-band traits (Goldsmith et al., 2000). Indeed, extraversion-related traits exhibit the greatest relative stability in adulthood (Roberts & DelVecchio, 2000; Vaidya, Gray, Haig, & Watson, 2002). The child literature is less well developed in this regard, as most stability studies of child temperament have examined only a small number of traits.

Stability must be considered in light of two important issues: (a) normative developmental change in temperament and emotional functioning stemming from maturation or experience (Goldsmith et al., 2000; Rothbart & Bates, 1998) and (b) methodological
issues in the measurement of temperament. Relative instability would be expected for measures taken early in life, particularly for traits that rely on later maturing processes (Rothbart, Derryberry, & Hershey, 2000). Normative developmental change may also lower rank-order stability across periods of rapid developmental reorganization in the systems governing particular traits. With regard to emotional traits, there is considerable evidence that children’s experience and expression of emotions change in important ways from infancy through childhood (Abe & Izard, 1999; Thompson, 1988).

Stability estimates are also influenced by the measurement strategy used to tap child temperament. The bulk of the literature on stability of child temperament has relied on questionnaire methods, specifically parental reports. Indeed, this method has provided strong evidence for the stability of child temperament. For example, Pedlow and colleagues (1993) examined the stability of maternal reports of temperament from infancy to 8 years of age and found that measures of approach, irritability, cooperation/manageability, persistence, rhythmicity, and inflexibility all exhibited moderate to high temporal stability; however, stability appeared to be greater for intervals after infancy. Lemery, Goldsmith, Klimmt, and Mrazek (1999) collected maternal reports on multiple temperament questionnaires at 3, 6, 12, 18, 24, 36, and 48 months, including the traits of PE, distress–anger, fear, and activity level. These traits were generally moderately stable over time, and stability appeared to increase from infancy to the toddler-preschool age period. Similarly, Roberts and DelVecchio’s (2000) meta-analysis of personality stability across the life span found that trait stability was lowest (.35) for ages 0–3 and increased to .52 in the period between 3 and 6 years of age.

Parent report measures have considerable advantages, particularly their ecological validity. These measures capitalize on parents’ history of knowledge regarding their child’s behavior, can cover a broad range of traits, and are efficient and economical (Mangelsdorf, Schopp, & Buur, 2000; Rothbart & Bates, 1998). In addition, parent report measures have been developed that map onto modern theoretical models of temperament that emphasize continuity between child and adult personality (e.g., Rothbart, Ahadi, Hershey, & Fisher, 2001). However, parent report measures also have important limitations. First, they have low convergent validity with laboratory measures, naturalistic observations, and teacher reports (Kagan, 1998; Seifer, Sameroff, Barrett, & Krafftchuk, 1994) and have poorer predictive validity for later psychological adjustment compared with self- and teacher reports (Mesman & Koot, 2001). Second, they are subject to dysphoria-related biases (Youngstrom, Izard, & Ackerman, 1999). Third, parent report measures confound at least two processes that may exhibit differential true stability across development: child temperament and parental interpretations of child behavior. Parental interpretations may be influenced by factors other than the child’s behavior, such as parents’ desire to maintain a consistent view of their child (Kagan, 1998) or stable characteristics of the parent (Bates, 1994). To the extent that parents infer continuity of their child’s behavior over time or hold stable, but inaccurate, interpretations of their child’s personality, parent report measures may result in inflated stability estimates for child temperament traits.

Laboratory measures of child temperament have several unique advantages relative to parent report. Lab tasks that use standardized stimuli designed to elicit behaviors relevant to the traits of interest allow for a sharper differentiation of individual differences in child behavior from the contexts in which they are observed. This may be particularly important for measuring emotional reactivity. Behavior during lab tasks can be coded using objective criteria, avoiding the problem of parental interpretation of child behavior. Naturalistic observations also circumvent this problem, but do not control for the influence of context on child behavior. Laboratory assessments also make it possible to expose children to contexts that may elicit behaviors that are expressed at a lower base rate in naturalistic settings (such as fear). However, lab measures also have notable drawbacks, including the considerable investment of time and expense necessary, possibly lower ecological validity, and the challenge of designing measures that are equally valid for tapping temperament across developmental periods. Thus, although laboratory measures provide an important means of addressing the stability of child temperament and may yield a more stringent test of stability than parent report questionnaires, stability estimates derived from laboratory measures should be compared with those derived from other methods.

Only a small number of studies have reported on the stability of child temperament as measured by laboratory tasks; the majority of these involved infant or toddler participants and examined relatively brief intervals between assessments. Goldsmith and Campos (1990) found moderate stability across a 2-week interval for lab measures of fear and pleasure (.38 and .55, respectively) in 9-month old twins. Carniiero, Perez-Lopez, Gonzalez-Salinas, and Martinez-Fuentes (2000) examined the stability of lab tasks tapping emotional tone (a bipolar scale of negative to positive affect), activity level, and sociability across ages 3, 6, 9, and 12 months. Stability across the various intervals ranged from low to moderate and tended to be higher at the older ages (9 to 12 months), at which point all traits but activity level demonstrated moderate (and comparable) stability. Rothbart, Derryberry, and Hershey (2000) reported on stability of lab measures of smiling and laughter, fear, frustration, and distress collected at 3, 6, 10, and 13.5 months of age. Stability estimates ranged from low to moderate. They also reported on the correlation between lab measures of these traits at 13.5 months and maternal reports of child temperament collected at 7 years of age; there was some evidence for stability of corresponding traits and continuity between traits thought to reflect the same temperament superfactors. Belsky, Hsieh, and Crnic (1996) examined stability of positive and negative emotional expressions collected at 12–13 months and at 18–20 months. NE at 18–20 months was strongly predicted by earlier NE (r = .67), but not PE; similarly, PE at 18–20 months was predicted by earlier PE (.79), but not NE. Thus, this study yielded evidence for both the stability of these two traits and for their relative independence.

Numerous studies have reported on the stability of laboratory-assessed behavioral inhibition, a trait marked by fearfulness, social reticence, and distress to novelty (see Fox, Henderson, Marshall, Nichols, & Ghera, 2005 for a review); these studies have indicated that behavioral inhibition is at least moderately stable across childhood and into adulthood (e.g., Gest, 1997). Pfeifer, Goldsmith, Davidson, and Rickman (2002) followed up a group of 7- to 8-year-old children who had been categorized as either extremely behaviorally inhibited, extremely uninhibited, or intermediate on behavioral inhibition at 32 months of age. Behavioral inhibition and exuberance (a composite of smiling, laughter, and approach behaviors) were coded from lab tasks conducted at ages 7–8. The two extreme groups
categorized as inhibited and uninhibited as toddlers differed significantly on the laboratory behavioral inhibition composite at ages 7–8; the previously uninhibited group exhibited higher exuberance in the laboratory at ages 7–8 than did the previously intermediate group.

In summary, the existing literature from both parent report and laboratory studies of child temperament traits provides support for the claim that temperament exhibits stability over time. However, there are limitations to conclusions that can be drawn from this literature. First, parent report questionnaires may overestimate stability to the extent that these measures tap stable, but inaccurate, parental interpretations. Second, studies of the stability of laboratory measures of temperament have typically examined a small number of traits, often those at the subordinate level. Third, most stability studies have focused on infancy, which now appears to be a developmental period characterized by lower temperament stability compared with the preschool period and beyond (Caspi, 2000). Finally, the majority of these studies have used relatively brief (less than 1 year) interassessment intervals.

The Current Study

Our aim in the current study was to explore the stability of child temperamental emotionality traits across the ages of 3 to 7 years, focusing on structured laboratory tasks designed to measure individual differences in two temperament superfactors, PE and NE, defined by emotional reactivity and their lower order facets. Following models focusing on the affective components of temperament (Rothbart, Ahadi, et al., 2000; Watson & Tellegen, 1985), we grouped lower order facets as follows: PE—positive affect, anticipatory positive affect, sociability, engagement, and activity level; NE—sadness, anger, and fear.

One hundred children completed a battery of laboratory tasks tapping temperament at age 3; 46 were reassessed with a separate battery of laboratory tasks at ages 5–6, and 64 completed another laboratory assessment at age 7. Parent reports of child temperament were collected at each assessment; naturalistic home observations of child temperament were conducted at age 3. We have previously reported on associations between our age 3 temperament measures and maternal depression (Durbin, Klein, Hayden, Buckley, & Moerk, 2005), child behavior problems (Hayden, Klein, & Durbin, 2005), and child cognitive styles (Hayden, Klein, Durbin, & Olino, 2006).

On the basis of the existing longitudinal literature on child temperament, we hypothesized that all traits would exhibit at least small to moderate effect sizes for stability across the longest time period (ages 3–7) and that stability coefficients would be higher for shorter interassessment intervals. We had no strong expectations about whether stability would increase with age (i.e., whether the ages 5–6 to age 7 interval would yield higher stability than the age 3 to ages 5–6 interval). On the basis of the adult literature on relative stability, we did expect differential stability across traits, specifically that extraversion (PE)-related traits would exhibit higher stability than neuroticism (NE)-related traits (Roberts & DelVecchio, 2000; Vaidya et al., 2002). Regarding differential stability of our affective facets (positive affect, sadness, anger, and fear) compared with more behaviorally defined traits (interest/engagement, sociability, and activity level), two competing predictions were possible. First, there is evidence from adult samples that self-report measures of affective state exhibit lower stability than personality traits (Vaidya et al., 2002), suggesting that positive and negative affect measures should be less stable than the behavioral components of extraversion (PE) and neuroticism (NE). However, some theorists have proposed that emotionality traits are more core biological aspects of personality (Goldsmith et al., 2000), implying that measures of positive and negative affect might exhibit higher stability than other temperament traits.

Method

Participants

Participants were 100 children from Long Island, New York. Children were initially recruited for a study of child temperament risk for depression (Durbin et al., 2005) using two methods: contacting families in local ZIP codes via a commercial mailing list (51.9%) and requesting volunteers for a study of children’s personality through ads in local newspapers and preschools (48.1%). Participants obtained through the two methods did not differ on any of the child temperament variables used in this study. Children were between the ages of 3 years, 0 months, and 3 years, 11 months. Consistent with 2000 census data for Suffolk County, New York, where the study took place, the sample was mostly White (85.0%), children came predominantly from two-parent homes (97%), and most of the mothers in the sample worked outside the home to some extent (58%). The sample was primarily working class or middle class, with a mean score of 34.9 ($SD = 9.9$) on Hollingshead’s (1975) Four Factor Index of Social Status. The mean age of parents was 33.6 years ($SD = 4.3$) for mothers and 36.4 years ($SD = 5.2$) for fathers. The mean age of child participants was 3.6 years ($SD = 0.3$), and 45% were female. Children were assessed for temperament traits using three methods: parent report, naturalistic home observations, and laboratory tasks. Children were reassessed at two later time points as part of a larger study tracing the link between low PE and markers of risk for depression. Approximately 2.25 years after the initial assessment (mean time between Time 1 and Time 2 [T1 and T2] = 2.22 years, $SD = 0.35$, range = 1.34–3.08), an extreme groups design was used to examine the association between low PE and laboratory measures of reinforcement and punishment sensitivity. Children who scored at the extreme ends of the distribution of positive affect (PA) scores at age 3 were invited back to the laboratory for a new battery of tasks. Twenty-three low-PA children were recruited from the 32 bottom-scoring children. Of the 9 remaining low-PA children, 2 families had moved out of the area, 3 refused to participate, and we were unable to contact 4. Twenty-three high-PA children were recruited from the 25 top-scoring children. We were not able to contact the remaining 2 families. We compared those targeted participants (lowest and highest PA at age 3) who completed the follow-up lab visit with those targeted participants who did not on the following variables: age 3 negative affect (NA) and family socioeconomic status (Hollingshead, 1975). Targeted children who did not participate in the follow-up did not differ from those who did participate on either of these variables. At T2, the 46 children ranged in age from 4 years, 11 months, to 6 years, 4 months, with a mean age of 5.74 years ($SD = 0.35$); 47.6% were female.

At age 7 (approximately 1.25 years after the T2 assessment, $M = 1.21$ years after T2, $SD = 0.43$, range = 0.30–2.21), a subset
of the original T1 sample was reassessed to examine the development of depressogenic cognitions (Hayden et al., 2006), using a different battery of lab tasks. One of the 100 children was used to pilot the follow-up tasks. Follow-up assessments were obtained for 64 of the remaining 99 children (64.6%; 32 boys and 32 girls). Of the 35 families not available for follow-up, 3 had moved, 23 declined to participate, and we were unable to reach 8. Families who declined to participate typically stated that they were too busy. At follow-up, mean child age was 7.0 years (SD = 0.5). Children who were and were not followed up did not differ on family socioeconomic status, age 3 PA, or age 3 NA. Thirty-seven of the child participants assessed at age 7 had also been assessed at ages 5–6.

Measures

Parent Reports of Temperament

We collected maternal reports at T1, T2, and T3 on the Child Behavior Questionnaire (CBQ; Rothbart et al., 2001). The CBQ is a widely used 195-item caregiver report measure of temperament for 3- to 7-year-old children; its scales have adequate internal consistency (Rothbart et al., 2001). For this study, we used the following scales: Smiling/Laughter, Shyness, Sadness, Anger, Fear, and Activity Level. The CBQ was completed by 99% of the mothers when the child was age 3, 72% when the child was ages 5–6, and 85.9% when the child was age 7.

T1 (Age 3) Laboratory Assessment of Temperament

At age 3, children completed 12 tasks designed to measure individual differences in PE- and NE-related traits. The laboratory battery consisted of portions of the Laboratory Temperament Assessment Battery (Goldsmith, Reilly, Lemery, Longley, & Prescott, 1993). Our use of the Laboratory Temperament Assessment Battery has been described in greater detail elsewhere (Durbin et al., 2005; Hayden et al., 2005). Child emotionality elicited by the Laboratory Temperament Assessment Battery has been linked to relevant constructs such as emotional tone during mother–child interactions (Kochanska, Coy, Tjebkes, & Husarek, 1998) and emotion regulation (Buss & Goldsmith, 1998). Tasks are described below in the order in which they were administered, along with the traits they were primarily intended to elicit.

Risk room (fear, activity level). In this task, the child was allowed to play freely by him- or herself with a set of novel and ambiguous stimuli (a cloth tunnel, short staircase, mattress, Halloween mask, balance beam, and a large, black cardboard box) in the main laboratory area. After 5 min, the experimenter returned and asked the child to approach each object.

Tower of patience (engagement). The child and experimenter took turns building a tower of cardboard blocks. The experimenter adhered to a schedule of delays of increasing length before placing her block on the tower (5, 10, 15, 20, and 30 s), forcing the child to wait longer each time to take his or her turn.

Arc of toys (PA, engagement, NA). The child was allowed to play freely by him- or herself in a roomful of toys. After 4 min, the experimenter returned to ask the child to clean up the play area by putting the toys in a box. Stranger approach (fear, NA). At the beginning of the task, the child was left alone in the main experimental area. After a few moments, a friendly male research assistant entered the room and spoke to the child while gradually walking closer.

Make that car go (PA, engagement, activity level). The experimenter and child raced with two remote-controlled racecars.

Transparent box (NA, engagement). The experimenter locked an appealing toy in a transparent box and left the child with a set of inoperable keys to use to open the box. After 3.5 min, the experimenter returned, explaining that she had accidentally given the child the wrong keys. The child was then allowed to unlock the box and play with the toy.

Pop-up snakes (PA, anticipatory PA). The experimenter showed the child what appeared to be a can of potato chips, actually containing coiled-spring snakes. The child was then encouraged to surprise his or her mother with the snakes.

Snack delay (impulsivity). The child was instructed to wait for the experimenter to ring a bell before eating a piece of candy. The experimenter waited a systematic series of delays before ringing the bell (10, 20, and 30 s).

Impossibly perfect green circles (NA). The experimenter asked the child to draw a circle on a piece of paper, then responded by mildly criticizing the child’s work and asking the child to draw another circle. This process was repeated for 2 min.

Popping bubbles (PA, engagement, activity level). The child and experimenter played together with a bubble-shooting toy.

Painting a picture (PA). The child played with watercolor pencils and crayons.

Box empty (anticipatory PA, NA). The child was given an elaborately wrapped empty box to open, under the pretense that an appealing toy was inside. After a brief interval in which the child was left alone to discover that the box was empty, the experimenter returned with several small toys for the child to keep, explaining that she had forgotten to place the toys inside.

Coding of laboratory episodes. Episodes were coded using a global system in which coders watched an entire episode and made a single rating for each variable based on all behaviors thought to be relevant to that trait. A common set of traits was rated across all episodes, regardless of the trait any particular task was designed to elicit. This approach has the advantage of capturing instances of normative behavior that were not specifically predicted by the task’s design, as well as increasing the number of ratings, allowing for more reliable estimates. We have previously reported (Durbin et al., 2005) that this global coding system exhibited high convergent validity with a microcoding system involving coding in discrete time intervals.

In the global coding system, discrete emotions (happiness/PA, anger, sadness, and fear) were assessed by separately coding and counting instances of facial, vocal, and bodily indicators of the emotion in each episode. Coding of facial expressions and their intensity used the AFFEX system (Izard, Dougherty, & Hembree, 1989). The total number of expressions at each intensity level (ambiguous/low intensity, definite/moderate intensity, definite/strong intensity) that occurred across the entire episode was considered when making a single, final rating for facial, vocal, and bodily indicators of each emotion. Total number of expressions at each intensity level were converted to scales ranging from 0 to 3 (vocal and bodily indicators) or 0 to 4 (facial indicators), with higher numbers indicating more frequent and intense expressions.
These ratings were averaged to produce composite variables for each emotion. For example, the PA scale reflects the frequency and intensity of smiling, positive verbalizations, and joyful bodily movements. Other traits (e.g., sociability and engagement) were coded on a single scale for each episode. Alphas (internal consistency of codes across episodes) and interrater reliabilities (intraclass correlation coefficients [ICCs]; assessed on a subsample of 15 cases) are reported below. The following scales were used in this report, each consisting of the average of codes across episodes:

- PA (α = .90; ICC = .94), sociability (α = .81; ICC = .93), engagement (the child’s level of interest in the tasks; α = .56; ICC = .72), anticipatory PA (PA/excitement occurring in anticipation of or before rewarding events; α = .58; ICC = .83), activity level (the vigor, energy, and extent of the child’s physical movements; α = .83; ICC = .75), NA (α = .77; ICC = .82), sadness (α = .67; ICC = .82), anger (α = .75; ICC = .84), and fear (α = .59; ICC = .66).

**T1 (Age 3) Home Observations**

Home observation ratings were made at the age 3 assessment using the Child Temperament and Behavior Q-Set (Buckley, Klein, Durbin, Hayden, & Moerk, 2002), which consists of 90 items comprising five scales and 11 subscales and was designed to tap temperamental and emotional behavior relevant to adjustment. Each child was observed during two independent home visits (mean length = 2 hr, 32 min, SD = 22 min) by trained pairs of graduate and undergraduate students who had no access to laboratory data. A different pair of raters observed the children at the second home visit, without knowledge of the data from the first home visit. After each observation, the pairs of raters independently sorted the 90 Child Temperament and Behavior Q-Set items into nine categories ranging from 1 (extremely uncharacteristic of the child) to 9 (extremely characteristic), producing a fixed distribution of scores for the child. Two independent sets of ratings were available from each home observation, yielding a total of four sets of ratings of each child’s behavior. Ratings were aggregated by averaging across all four raters. For the present study, we used the following scales: Affective Constriction (diminished positive affect), Low Appetitive Behavior (low pursuit of rewarding activities), Apathy (behavioral impoverishment), Low Sociability, Sadness/Depression, Anger, Fear of Novel Situations and Hyperactivity. Median interrater reliability, based on all four raters, was .79 (ranging from .75 for Anger to .86 for Fear of Novel Situations), and median internal consistency for these scales (coefficient alpha) was .83 (ranging from .78 for Fear of Novel Situations to .96 for Affective Constriction).

**T2 (Ages 5–6) Laboratory Tasks**

At age 5, children completed a new battery of laboratory tasks, some assessing behavior under rewarding and punishing conditions and others designed to be similar to those used at age 3. The 10 tasks are described below in the order in which they were conducted, along with the traits they were intended to tap.

- **Card sorting (PA, reward and punishment sensitivity).** Children were shown cards depicting geometric figures varying in color, number, and shape and were taught to sort the cards by color. Children sorted for several timed trials that varied by outcome (reward, noncontingent reward, punishment) and by the number of sorted cards required to obtain or avoid the contingency.

- **Jenga tower (NA).** The experimenter and child took turns removing pieces from a tower made of Jenga blocks. Each block was marked with a number; players scored the number of points on the block they removed. Children were told the player with the most points was the winner, but if either knocked over the tower, that player lost. The experimenter adhered to a series of delays before taking her turns to elicit frustration. Two trials were conducted. In Trial 1, the experimenter took only low-scoring blocks before knocking over the tower during the seventh turn. In Trial 2, the experimenter deliberately removed high-scoring blocks to elicit competitive behavior from the child participant.

- **Disappointing toy (NA, activity level).** The experimenter showed the child two pictures, one of a small plastic doll and the other of remote-controlled racers, and asked the child which she or he would prefer. The experimenter left the room and returned with the nonpreferred toy (generally the doll). After 1 min, an assistant entered with the preferred toy (usually the racers), and the child and experimenter played together for 3.5 min.

- **Can’t stop (PA, reward and punishment reactivity, engagement, anticipatory PA).** The child played three trials of a computerized game of luck in which he or she could win several small prizes that the child first selected and ranked according to their desirability. The experimenter placed the most preferred next to the highest point on a scoring track and the least preferred on the lowest point. In each trial, the child selected two “lucky colors” and used a computer mouse to “spin” two displays that revealed two different colored circles on the monitor. Each color was represented by a marker on the scoring track which was advanced to a higher prize level each time the child spun to reveal his or her lucky color. If the color did not appear, the corresponding marker was removed and no prize was given. After each successful spin, the child was given the choice to settle for the prizes at the current level of his or her marker or to risk losing them in order to take a chance at advancing to more highly desired prizes. The game was predetermined such that children who chanced at least three spins on each trail were guaranteed to win four prizes. All children were given their most highly ranked prize at the end, regardless of their performance.

- **Mixed-up puzzle (NA, engagement).** The experimenter asked the child to assemble a puzzle she described as “really easy, for kids much younger than you.” The child was given pieces from two highly similar but different puzzles, making it impossible to complete. After 3 min, the experimenter returned and explained that the puzzle was impossible to finish.

- **Practical joke (PA, sociability, anticipatory PA).** The experimenter showed the child how to operate a remote controlled whoopee cushion that emitted two sounds when a button on the remote was pushed. The experimenter asked the child if she or he would like to trick his or her parent with the joke and then showed the child how to hide the whoopee cushion under a chair. The parent was then invited into the room, and the child tricked the parent when she or he sat on the chair.

- **Story time (fear, sociability).** The child was asked to tell a story to an unfamiliar research assistant, whom the experimenter described as “a story expert.” The child was given the picture book *A Boy, a Dog, and a Frog* by Mercer Mayer and asked to use the book to tell the story to the assistant, who would then give the child...
a grade on how well she or he told the story. After 4 min, the experimenter returned, and the assistant praised the child’s story.

Puzzles with parent (PA, sociability, engagement). The parent was taught how to play a puzzle game (Rush Hour) that involved setting up toy cars on a grid as depicted on item cards, then moving the cars to move a single target car out of the grid. The mother was asked to teach the child how to play, and the pair worked on as many puzzles as possible in 5 min.

Toy parade (PA, activity level, engagement). The child was allowed to play with a series of toys, one at a time. The child was given a bell that she or he could ring to indicate that she or he wanted to trade the current toy for the next, but were not told the identity of the new toy. When the child rang the bell, a research assistant removed the current toy and replaced it with a new one. Toys appeared in the following order: Mr. Potato Head, fun hop (an inflatable ball the child sits and hops on), Gearation (a magnetized gear construction toy), an electronic guitar and a floor piano, a large Lego set, a bubble toy, and a small minigolf set.

Coding of age 5 tasks. Lab tasks were coded using the same global method as for the age 3 tasks. PA, anticipatory PA, sociability, engagement, activity level, and NA (sadness, anger, and fear) were coded across all 10 tasks and averaged. Alphas (internal consistency of codes across episodes) and interrater reliabilities (ICCs, assessed on a subset of 15 cases) were as follows: PA, $\alpha = .93$, ICC = .97; anticipatory PA, $\alpha = .63$, ICC = .53; sociability, $\alpha = .86$, ICC = .91; engagement, $\alpha = .67$, ICC = .87; activity level, $\alpha = .87$, ICC = .97; NA, $\alpha = .84$, ICC = .91; sadness, $\alpha = .69$, ICC = .81; anger, $\alpha = .67$, ICC = .90; and fear, $\alpha = .47$, ICC = .84.

T3 (Age 7) Laboratory Tasks

At age 7, children completed three tasks designed to measure depressogenic cognitions and instrumental and social helplessness (Hayden et al., 2006). Although not specifically designed to elicit individual differences in temperamental emotionality, the tasks were similar to the age 3 and ages 5–6 laboratory tasks in that they involved interacting with an experimenter, presented children with challenging tasks, and, most important, were emotionally evocative. Furthermore, we used the same coding methods as those for the age 3 and ages 5–6 laboratory tasks for measuring temperaments in these tasks. The age 7 tasks are described below in the order in which they were administered.

Instrumental helplessness. Following Nolen-Hoeksema, Wolfson, Mumme, and Guskin (1995), children were given a set of blocks and asked to reproduce designs on printed cards similar to those found in the Wechsler Intelligence Scale for Children—Revised (Wechsler, 1974) Block Design subtest. The first, third, and fifth puzzles were solvable, and the second, fourth, and sixth were not. Children were given 4 min to work on the solvable puzzles to ensure that they had sufficient time to complete them, thus creating the expectation that all the puzzles were solvable. Children were given 2 min to work on the unsolvable puzzles to limit frustration. After the last puzzle, children were allowed to select one of the puzzles for further play.

Interpersonal helplessness. Children were told they were going to be interviewed for admission to a club “just for kids” (Erdley, Cain, Loomis, Dumas-Hines, & Dweck, 1997). An unfamiliar interviewer asked children a series of questions under the pretense that she needed to determine whether they would get along well with the club members. The interviewer told the child she was going to send his or her information to the club president by computer, so that the president could immediately decide whether the child would be admitted. After a brief delay, the assistant returned, stating that the president was not sure about whether the child should be admitted and that she or he wanted to know more about the child before deciding. Children were allowed to choose whether they wanted to reapply to the club by providing more information about themselves. Children were then asked several questions to assess their attributions for the ambiguous rejection. The episode ended when the assistant returned with a certificate of membership, explaining that she had made a mistake and that the child had been accepted into the club from the very beginning.

Perfectionism. Children were invited to participate in a contest to see who could draw the best copy of a poster presented to them. They were given 4 min to use markers and paper to produce a copy. On completion, children were asked what they did and did not like about their picture and whether they thought their picture could win the contest. The experimenter invited a research assistant “contest judge” to evaluate the child’s picture; the judge proceeded to point out some discrepancy between the child’s copy and the original poster. Children were then asked to make an overall rating of their copy and were offered the opportunity to see another child’s work, either “someone who did much better than you or someone who did much worse than you.” Finally, the research assistant judge awarded the child a first-place ribbon for the contest.

Coding of age 7 laboratory tasks. Coding was done using the same global system as for the age 3 and ages 5–6 tasks. Interrater reliability was assessed on a subsample of 29 participants. The following traits were coded: PA (\(\alpha = .94\), ICC = .90), sociability (\(\alpha = .94\), ICC = .90), engagement (\(\alpha = .74\), ICC = .57), NA (\(\alpha = .94\), ICC = .89), sadness (\(\alpha = .93\), ICC = .86), and anger (\(\alpha = .86\), ICC = .74). Fear/anxiety was coded only in the interpersonal helplessness task (ICC = .66). Anticipatory PA and activity were not coded, as none of the age 7 tasks were designed to elicit individual differences in these traits.

Analyses

Below, we report on (a) the relative stability of temperament traits measured by laboratory tasks and home observations at age 3 and laboratory tasks at ages 5–6 and age 7; (b) relative stability of parent report measures across ages 3, 5–6, and 7; (c) the heterotypic continuity of the facets of PE (positive affect, sociability, engagement, and activity level) and NE (sadness, anger, and fear) over time; and (d) the independence of PE and NE across ages 3, 5–6, and 7.

Results

Relative Stability of Temperament Traits Across Ages 3, 5–6, and 7

The following traits were examined: PA (composite of facial, vocal, and bodily indicators of positive affect in the laboratory; ratings of affective constriction at the age 3 home observation
(reversed); the Smiling/Laughter scale from the CBQ; anticipatory PA (anticipatory PA in the lab; appetitive behavior at the age 3 home observation); sociability (sociability in the lab; Sociability scale from the age 3 home observation; CBQ Shyness scale [reversed]); engagement (engagement from the lab; behavioral apathy ratings from the age 3 home observation [reversed]); activity level (activity from the lab; activity ratings from the age 3 home observation; CBQ Activity scale); NA (composite of facial, vocal, and bodily indicators of sadness, anger, and fear from the laboratory); sadness (composite of facial and bodily sadness from the lab; sadness/depression ratings from the age 3 home observation; CBQ Sadness scale); anger (composite of facial and bodily indicators of anger from the lab; anger ratings from the age 3 home observation; CBQ Anger scale); and fear (composite of facial and bodily fear from the lab; fear of novel situations ratings from the age 3 home observation; CBQ Fear scale).

Laboratory Measures

All temperament traits showed some stability across each of the three intervals (age 3 to ages 5–6, ages 5–6 to age 7, and age 3 to age 7). Correlation coefficients for each trait across all time intervals are shown in Table 1. Unsurprisingly, stability coefficients were highest for the shortest interval (ages 5–6 to age 7) and lowest for the longest interval (age 3 to age 7). Collapsing across facets of each superfactor, stability was similar for PE- and NE-related traits, with the exception of the ages 5–6 to age 7 interval. For that interval, the median r was .60 for PE traits compared with .40 for NE traits. By contrast, the median r was .50 for PE traits and .44 for NE traits for the ages 3 to ages 5–6 interval and .36 for PE traits and .27 for NE traits for the ages 3 to age 7 interval. The higher coefficients for PE traits in the ages 5–6 to age 7 comparison may have been due to the selection criteria at ages 5–6 (the top and bottom quartile of age 3 PA scores), which resulted in more extreme groups on PA at this assessment. Owing to this feature of the design, our estimates of PE stability involving the ages 5–6 interval should be interpreted with caution.

Stability coefficients were compared by using Fisher’s r-to-z transformation and testing whether the difference between the resulting z scores was statistically significant. Across each time interval, emotional expression variables (PA, NA, sadness, anger, and fear) had nonsignificantly larger stability coefficients than other traits (with only two exceptions). Individual negative emotions (sadness, anger, and fear) also exhibited differential stability over time. Fear was the least stable across all intervals, and sadness was more stable than anger at all intervals with the exception of that between ages 3 and 7. Although none of these stability coefficients were significantly different from one another (our power to detect differences was low in this small sample), the magnitude of difference between the stability of fear and the other two negative emotions (anger and sadness) ranged from .09 to .31. The stability of fear across intervals involving the age 7 assessment may have been deflated due to the nature of the task used at that time point, which was more likely to tap social anxiety than fear.

Home Observation Measures

We also examined the predictive validity of age 3 home observation variables for lab measures of traits at ages 5–6 and 7. Our home observation measures exhibited moderate convergence with concurrent lab measures of the same traits at age 3 (median r = .29), with correlations ranging from a low of .03 (home and lab sadness) to a high of .52 (lab and home PA). Stability estimates for the age 3 to ages 5–6 and age 3 to age 7 intervals are shown in Table 1. For all but one trait (sadness), stability coefficients were larger for the age 3 to ages 5–6 interval than for the age 3 to age 7 interval. In general, the stability of PE traits was higher than that for NE traits. The associations between age 3 home observation measures and ages 5–6 laboratory measures were smaller than the

Table 1

<table>
<thead>
<tr>
<th>Trait</th>
<th>Baseline and follow-up lab measures</th>
<th>Baseline home observation → follow-up lab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age 3 → Ages 5–6 (n = 46)</td>
<td>Ages 5–6 → Age 7 (n = 37)</td>
</tr>
<tr>
<td>PE/extraversion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>.59 (.62) *</td>
<td>.70 (.73) *</td>
</tr>
<tr>
<td>Anticipatory PA</td>
<td>.41 (.69) **</td>
<td>—</td>
</tr>
<tr>
<td>Sociability</td>
<td>.52 (.58) ***</td>
<td>.62 (.69) ***</td>
</tr>
<tr>
<td>Interest/engagement</td>
<td>.37 (.40) ***</td>
<td>.48 (.68) ***</td>
</tr>
<tr>
<td>Activity level</td>
<td>.62 (.73) *</td>
<td>—</td>
</tr>
<tr>
<td>NE/neuroticism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA composite</td>
<td>.57 (.66) *</td>
<td>.59 (.66) *</td>
</tr>
<tr>
<td>Sadness</td>
<td>.52 (.64) *</td>
<td>.52 (.62) **</td>
</tr>
<tr>
<td>Anger</td>
<td>.40 (.46) **</td>
<td>.30 (.37) ****</td>
</tr>
<tr>
<td>Fear</td>
<td>.23 (.31) ****</td>
<td>.21 (.28)</td>
</tr>
</tbody>
</table>

Note. Indices in parentheses are corrected for attenuation owing to interrater unreliability. Sadness, anger, and fear variables include facial and bodily indicators of the emotion. Negative emotionality (NE) variables are a composite of facial and bodily indicators of each emotional state and negative verbalizations. Positive emotionality (PE) variables are a composite of facial, vocal, and bodily positive affect. Home observation scales: PA = affective constriction; Anticipatory PA = appetitive behavior; sociability = sociability; engagement = apathy; activity Level = hyperactive; NA = composite of sadness, anger, and fear; Sadness = sadness/depression; anger = anger; fear = fear of novel situations. Dashes indicate that measures of that trait were not made at one or more of the assessments.

* p < .10. ** p < .05. *** p < .01. **** p < .005. ***** p < .0001.
corresponding stability coefficients for laboratory measures for all but one trait (engagement); however, the magnitude of these differences was small for PE traits. In general, despite methodological differences between the laboratory and naturalistic home observations, PE traits exhibited moderate stability across these two methods at intervals of 2 and 4 years. By contrast, NE traits exhibited little to modest stability across these methods at 2- and 4-year intervals.

**Stability of Observational Measures, Corrected for Attenuation**

We also computed stability coefficients corrected for attenuation due to interrater unreliability. Although nearly all of our lab and home observation measures had adequate reliability, there was variability across constructs on the extent to which they were measured reliably (in particular, NE traits had lower reliability than PE traits). Corrected stability coefficients are shown in parentheses in Table 1. In general, stability estimates for temperament traits ranged from moderate to high across the two shorter (2-year and 1.25-year) intervals and from low to moderate across the 4-year interval.

**Parent Report Measures**

Forty-two mothers completed the CBQ at both the age 3 and ages 5–6 assessments, 55 completed the measure at both ages 3 and 7, and 55 at both ages 5–6 and 7. We chose the following CBQ scales to correspond to our laboratory measures of temperament: Smiling/Laughter (PA), Sadness (sadness), Anger (anger), Fear (fear), Shyness (sociability), and Activity Level (activity level). Stability coefficients are shown in Table 2. All scales exhibited at least moderate stability across each interval. Stability was slightly lower for the longest interval (age 3 to age 7; mean \( r = .60 \)) than for the two shorter intervals (mean \( r = .68 \) for both the age 3 to ages 5–6 interval and the ages 5–6 to age 7 interval). Similar to findings for the stability of lab assessments, fear was the least stable scale for two intervals and the second lowest in stability for the remaining interval. Scales tapping PE traits (Smiling/Laughter, Sociability, and Activity Level) were slightly more stable than those assessing NE (Sadness, Anger, Fear). In general, stability coefficients were higher for maternal reports than for lab measures (mean \( r = .60 \) vs. .31 from ages 3 to 7; mean \( r = .68 \) vs. .49 for ages 5–6 to 7; mean \( r = .68 \) vs. .47 from ages 3 to 5–6).

We also examined the cross-time correlations between our lab and parent report measures (earlier lab measures to later parent report measures, and vice versa). We have previously reported modest to moderate convergence between these two methods at age 3 in this sample (Hayden et al., 2005). Comparable levels of convergence were evident at the ages 5–6 and age 7 assessments. Similar to the findings linking age 3 home observation variables to later lab measures, associations across method and time for lab and parent report variables ranged from low (—.08 for age 3 parent report of anger to age 7 laboratory anger) to moderate (.50 for ages 5–6 laboratory fear to age 7 parent report of fear). The mean association between earlier lab measures and later parent report variables was .22, and the corresponding value from earlier parent report to later lab measures was .18. Thus, similar to Rothbart, Derryberry, et al. (2000), we found some evidence for continuity between CBQ scales and laboratory measures.

**Heterotypic Continuity of Subtraits of PE and NE**

Modern personality models emphasize the higher order structure of personality, with the superfactor of PE composed of lower order traits of positive affect, approach/reward-seeking behavior, sociability, dominance, warmth, and activity (Watson & Clark, 1997; Depue & Collins, 1999), and NE composed of negative affect, aggression/hostility, and negativistic thinking (Clark & Watson, 1999). Consistent with the proposed latent structure of these constructs, one would expect (a) that lower order traits marking the same higher order factor would exhibit continuity between one another over time, despite phenotypic differences among the narrow-band traits and (b) that higher order PE and NE factors would exhibit temporal stability.

To address the first issue, we computed cross-time correlations between PE and NE subtrait measures from the age 3 laboratory visit and the two later assessments; these are shown in Table 3. Continuity coefficients for PE traits (correlations between one subtrait of PE and its other subtraits) ranged from low to moderate and, in most cases, were quite similar in magnitude to the stability coefficients for each subtrait. For NE, continuity was more variable. In particular, fear did not exhibit strong continuity with other facets of NE. High sadness and anger at T1 predicted lower levels

---

**Table 2**

**Temporal Stability of Maternal Report Measures of Temperament Traits on the Child Behavior Questionnaire From Age 3 to Age 7**

<table>
<thead>
<tr>
<th>Traits</th>
<th>Age 3 ( \rightarrow ) Ages 5–6 ( n = 42 )</th>
<th>Ages 5–6 ( \rightarrow ) Age 7 ( n = 55 )</th>
<th>Age 3 ( \rightarrow ) Age 7 ( n = 55 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive emotionality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smiling/laughter</td>
<td>.67</td>
<td>.76</td>
<td>.70</td>
</tr>
<tr>
<td>Shyness (sociability)</td>
<td>.74</td>
<td>.72</td>
<td>.61</td>
</tr>
<tr>
<td>Activity level</td>
<td>.71</td>
<td>.79</td>
<td>.60</td>
</tr>
<tr>
<td>Negative emotionality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>.64</td>
<td>.56</td>
<td>.57</td>
</tr>
<tr>
<td>Anger</td>
<td>.72</td>
<td>.71</td>
<td>.66</td>
</tr>
<tr>
<td>Fear</td>
<td>.60</td>
<td>.57</td>
<td>.47</td>
</tr>
</tbody>
</table>

*Note. All ps < .0001.*
of fear at T3, and fear at T1 was largely unrelated to later NE traits. Similar results were found for the T2 to T3 interval (data are available on request). We also conducted simultaneous multiple regression analyses predicting later traits from earlier homotypic and heterotypic traits. Results indicated that in most analyses (13 of 20), the homotypic trait had the strongest relationship to the dependent variable among all the trait-subtrait predictors, but all analyses revealed significant associations among different traits across time, consistent with the correlational data.

Given that activity level has been linked to both NE and PE in child samples (Rothbart & Bates, 1998), we also examined conti-


tuity between activity and NE traits. The correlations between age 3 activity level and NE traits at ages 5–6 and age 7 were uniformly lower than those between age 3 activity level and PE traits at the two later assessments (mean \( r = .21 \) vs. .40). Similar results were found between age 3 PE and NE traits and ages 5–6 activity level. Therefore, our findings suggest that activity level may be a better marker of PE than NE in preschool and early childhood.

Regarding the second issue, our sample size was not sufficient to perform structural equation modeling of the stability of latent PE and NE traits. However, we did examine the stability of PE and NE composites, calculated as the mean of the five PE subtraits and of the three NE subtraits at each assessment. Unsurprisingly, these composite scales generally exhibited greater stability than did the subtrait scales, ranging from moderate to high. Stability was greatest for the shortest interval between T2 and T3 (68 for PE vs. .53 for NE) and somewhat lower between T1 and T2 (.57 for PE vs. .49 for NE) and between T1 and T3 (.46 for PE vs. .31 for NE).

**Independence of PA and NA**

Consistent with the literature on child temperament (e.g., Belsky et al., 1996) and adult personality and affectivity (Clark & Watson, 1999), we have previously reported that our T1 lab measures of PA and NA were orthogonal (Durbin et al., 2005). Examination of the cross-time correlations between lab measures of positive and negative emotions yielded further evidence for the independence of these constructs. Predicting T2 and T3 NA from earlier PA scores (T1 and T2) yielded a mean correlation of .12; the mean correlation between earlier measures of NA and later measures of PA was .15. The intercorrelation between PA and NA at the later assessments was higher than the cross-time–cross-trait associations and higher than for our T1 measures; PA and NA correlated .33 at both T2 and T3 (ps = .025 and .008 for T2 and T3, respectively) versus \( -.07 \) at age 3. The difference in intercorrelation between the initial and follow-up assessments could not be attributed to the subset of children assessed at T2 and T3 differing from the larger sample in the association between PA and NA at T1. For the subset of children who were followed up at T2, the intercorrelation between T1 PA and NA was .06; for those followed up at T3, it was \( -.01 \).

Discussion

We found considerable evidence for the stability of temperament traits across the ages of 3 to 7 in a community sample of children. Relative stability for the traits of PA, appetitive activity level, sociability, activity level, NA, sadness, and anger ranged from moderate to high across each of the three intervals (age 3 to ages 5–6, ages 5–6 to age 7, and age 3 to age 7), although stability for fear was lower. Our estimates were comparable to those reported for a similar age span in Roberts and DelVecchio’s (2000) meta-analysis and were consistent with the few studies that have reported on the stability of lab measures of temperament in infants and toddlers (Belsky et al., 1996; Carnicero et al., 2000). The stability of lab measures was impressive, considering that the tasks used to measure these traits differed at each age. Although not significantly higher in this relatively small sample, stability coefficients were larger for the emotional expression scales (PA and NA) compared with the other traits examined and lower for fear relative to other emotional traits (sadness, anger, positive

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuity Between Laboratory Measures of Subtraits of Positive Emotionality and Negative Emotionality Across Ages 3 (T1), 5–6 (T2), and 7 (T3)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>PA</td>
</tr>
<tr>
<td>Anticipatory PA</td>
</tr>
<tr>
<td>Sociability</td>
</tr>
<tr>
<td>Engagement</td>
</tr>
<tr>
<td>Activity level</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Sadness</td>
</tr>
<tr>
<td>Anger</td>
</tr>
<tr>
<td>Fear</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>T1</td>
</tr>
</tbody>
</table>
| Note. For comparison purposes, stability coefficients are shown in italics. T1 = age 3; T2 = ages 5–6; T3 = age 7. \( p < .10 \), \( ** p < .05 \), \( *** p < .01 \), \( **** p < .005 \), \( ***** p < .0001 \).
affect). This contrasts with the findings of Vaidya et al. (2002), who found that self-reported affective state was less stable than personality traits in an adult sample and is consistent with claims that emotional/affective traits are more “core” components of temperament (Goldsmith et al., 2000).

When stability coefficients were corrected for attenuation due to interrater unreliability, they approached the stability others have observed for parent report measures of temperament. For instance, Lamb, Chuang, Wessels, Broberg, and Hwang (2002) examined maternal ratings of child temperament using a Q-sort procedure (on the CCQ) across a similar time frame explored in our study and found stability coefficients of .50 for extraversion, .37 for agreeableness, .54 for conscientiousness, .38 for neuroticism, and .45 for openness to experience.

In addition to the considerable stability of lab measures, we also found stability across distinct methods of measurement. Home observation measures at age 3 were significantly related to lab measures of corresponding traits at ages 5–6 and 7. In most instances, the stability from home observation to later lab measures was similar to that from earlier to later lab measures. This was particularly true for PE traits, but less so for NE traits. We also observed stability across lab and parent report measures at different time points, although these correlations were somewhat weaker than those across home and lab measures. Convergence across multiple methods bolsters the claim that PE and NE traits exhibit stability across development and indicates that despite only moderate cross-method convergence at a single assessment, separate methods appear to each be tapping the stable components of PE and NE traits. This suggests that weak convergence between different temperament measures should not be automatically taken as evidence of poor construct validity. Future studies using structural equation modeling with larger samples could be used to explore the relative contribution of different methods for tapping stable elements of temperament, as well as to test questions regarding the invariance of the factor structure of temperament measures across development.

Consistent with the claim that temperament traits exhibit continuity across development (Caspi, 2000) and with the presumption that PE and NE are higher order temperament superfactors (Rothbart & Bates, 1998; Tellegen, 1985), we found that the lower order facets of each superfactor exhibited continuity with one another over time. PE subtraits were intercorrelated across intervals, often at a level as high as that for the stability of individual subtraits, supporting the claim that the facets of behavioral engagement, PA, sociability, and activity are part of a higher order PE superfactor. NE subtraits exhibited weaker continuity than PE facets. In particular, fear (which also exhibited low stability) was weakly related to sadness and anger at later time points. One possible explanation of this finding is that perhaps our lab tasks were more successful at tapping sadness and anger than fear. These weak associations might also be due to the poor internal consistency and interrater reliability for our lab measures of fear. A number of studies of behavioral inhibition, which includes elements of fear expression in addition to lack of approach to unfamiliar objects and people and social inhibition, have demonstrated considerable stability for this trait (Fox et al., 2005; Gest, 1997). Sadness and anger demonstrated stronger continuity with one another over time, possibly because both reflect an underlying emotional system responsive to the blockage of important goals (Lewis & Ramsay, 2005).

Finally, both child and adult models of temperament propose that PE and NE are orthogonal to one another (Clark & Watson, 1999; Rothbart & Bates, 1998). Consistent with this claim, we found that PE and NE were weakly associated with one another at the same assessment points. In addition, these traits did not predict one another across time. Of note, PE and NE appeared to be more strongly intercorrelated at the ages 5–6 and 7 assessments, compared with the baseline assessment. This may suggest real developmental change in the relationship between the two systems or subcomponents of these systems, or it could be due to differences in the tasks used at our various assessments. Further studies addressing this issue are necessary to replicate this finding and explore possible explanations.

In general, our results support the validity of models that emphasize PE and NE as temperament superfactors (e.g., Rothbart & Bates, 1998). However, other models emphasizing the distinction between approach- and withdrawal-related emotions and motivation (e.g., Davidson, 2001) might also predict stability of individual differences in particular emotions over time, but different patterns of continuity and intercorrelation among emotions (i.e., grouping anger and happiness as approach vs. fear and sadness as withdrawal emotions; e.g., Harmon-Jones & Allen, 1998). Although our results do not speak directly to this issue, weaker continuity for our NE measures suggests that alternative conceptualizations might explain patterns of individual differences in negative emotions over time. Approach–withdrawal models have generally been applied to describe emotional reactions to particular categories of contexts/stimuli, but have not been adapted to suggest alternative conceptualizations of temperament/personality, models of which have traditionally supported a distinction between traits related to positive emotions versus those associated with negative emotions (Clark & Watson, 1999).

Maternal reports of child temperament traits exhibited moderate to high stability across all time intervals in this sample. In this respect, our findings are quite similar to those in the literature (e.g., Pedlow et al., 1993). Stability coefficients were higher for maternal reports than for lab measures in our sample. A number of explanations for this finding are possible. First, questionnaire items were the same across assessments, whereas different lab tasks were used at each age; therefore, lab stability may have been lower due to differences in the tasks used. Second, parent report measures may be influenced by other stable processes in addition to child behavior (such as parental attitudes toward their child). It is important to note that coding of lab measures was done by separate raters at each time point, whereas parent reports were completed by the same individual at each time point. Finally, it is possible that parent report may more adequately tap the stable elements of child temperament.

This study had a number of considerable strengths. First, it included three assessment points separated by relatively lengthy time intervals, allowing us to examine stability across a longer interval than many existing studies in the literature. Second, multiple methods of measurement were examined in stability analyses, including structured lab tasks, naturalistic home observations, and maternal reports. Third, very few studies have reported on the stability of laboratory measures of temperament, despite the advantages of this method. Finally, we explored both homotypic and heterotypic continuity of traits.
However, this study also had notable weaknesses. First, the sample size was relatively small, particularly for intervals involving the ages 5–6 assessment. Thus, replication of our findings in a larger sample is necessary. Second, there was attrition by the age 7 assessment (35.4% of the original sample did not complete the T3 lab visit). However, attrition was not related to extreme scores on temperament traits at T1. Third, our lab measures of temperament necessarily differed at each age. To the extent that the tasks used at different ages were not equivalent, some factors related to tapping particular traits (such as fear), they may have produced underestimates of the stability of those traits. Similarly, low interrater reliabilities for some traits (e.g., engagement and fear at T3) likely also contributed to underestimation of stability coefficients.

Finally, an extreme groups design (low and high age 3 PA) was used to select participants for the ages 5–6 assessment; therefore, it is not known whether our stability estimates for PE-related traits from this interval are a true estimate of the degree of stability for PE in an unselected sample composed of children across the full range of PE scores. To address this issue, we compared (a) the stability of PE traits from T1 to T3 in the subsample followed up at T2 (extremes on T1 PE; \( n = 40 \) for the T1–T3 interval) to those not followed up at T2 (the middle of the T1 PE distribution; \( n = 24 \)) and (b) T1 to T3 stability for the subsample followed up at T2 versus that for the entire sample available for the T1–T3 interval. Each approach has a different disadvantage; the first is hampered by restriction of range in the middle group, which affects the magnitude of stability correlations, and the second contrasts stability in two groups with overlapping participants. Using the first strategy, stability was greater for the extreme groups compared with the middle group, but only for PE measures. The stability of lab measures of PA and sociability was higher in the extreme groups than in the middle PE group, whereas the reverse was true for engagement (mean \( r = .44 \) and \( .10 \) across PE traits for the extreme groups and the middle group, respectively). However, stability was similar for the extreme and middle groups on maternal reports of smiling, shyness, and activity level (mean \( r = .60 \) and .69, respectively), as well as for the T1 home observation to T3 laboratory tasks stability (mean \( r = .49 \) vs. .37). The stability estimates for laboratory-assessed NE were similar across the extreme groups followed up at T2 and the middle group (mean \( r = .22 \) vs. .30). The second strategy produced smaller differences between the extreme groups and the entire sample (mean \( r \) across PE traits = .44 vs. .36). Thus, our stability estimates for laboratory-assessed PE traits may have been inflated for the intervals involving the T2 assessment; replication in a larger sample following children across the range of PE is necessary to make conclusions about the stability of PE across this developmental interval.

The stability of temperament traits evident in this study could derive from a number of sources, including the action of genetic influences on temperament, the effect of stable characteristics of the environment in young children, or gene–environment interactions that contribute to the stability of individual differences in temperament traits. It is also possible that parenting or aspects of the parent–child relationship contributed to stability of child temperament traits. Several studies of behavioral inhibition have shown that behavioral inhibition stability is influenced by parenting behaviors that reinforce or limit behavioral inhibition in children (e.g., Rubin, Burgess, & Hastings, 2002). Given that stability was observed in this sample for observational measures in addition to parent report, factors other than stable parent perceptions of child temperament are implicated. Of note, measures of children’s frequency and intensity of expressions of happiness, sadness, and anger were somewhat more stable than the more behavioral traits, such as sociability and activity level. Thus, although children’s ability to regulate their expressions of emotion develops with age (Cole, Martin, & Dennis, 2004), individual differences in emotional expression exhibited relative stability over a considerable developmental interval (ages 3 to 7).

References


