The stability of temperament from early childhood to early adolescence: A multi-method, multi-informant examination

Daniel C. Kopala-Sibley[1], **Thomas Olino**[2], **Emily Durbin**[3], **Margaret W. Dyson**[4], and **Daniel N. Klein**[1]

[1]Department of Psychology, Stony Brook University, Stony Brook, NY

[2]Department of Psychology, Temple University, Philadelphia, PA

[3]Department of Psychology, Michigan State University, East Lansing, MI

[4]Child Mind Institute, New York, NY

Abstract

Temperament is a core aspect of children's psychological functioning and is assumed to be at least somewhat stable across childhood. However, little research has assessed the stability of temperament from early childhood to early adolescence. Moreover, few studies have examined the influence of measurement and analytic methods on the stability of early temperament over periods of more than a few years. We obtained laboratory observations and mother and father reports of temperamental negative and positive emotionality and effortful control from 559 three-year olds. Approximately 9 years later, children and both parents completed questionnaire measures of similar temperament constructs. Zero-order correlations revealed greater within- than cross-informant stability. In addition, compared to parent reports, early childhood laboratory measures showed greater convergent and divergent validity with child, mother, and father reports at age 12. Finally, latent temperament variables at age 3 composed of laboratory and parent-report measures and latent variables at age 12 composed of parent and child reports showed moderate stability. There was also a weak but significant association of early effortful control with later negative and positive emotionality. Results have implications for assessing temperament and knowledge of the stability of temperament across childhood.

Temperament, or relatively stable individual differences in emotional reactivity and regulation, has long been recognized as a core component of children's psychological makeup (Rothbart, Ahadi, & Evans, 2000; Zentner & Bates, 2008). It is also a key determinant of children's psychosocial functioning including psychopathology (Klein et al., 2012; Nigg, 2006), academic performance (Martin, 1994), and peer relationships (Sanson et al., 2004). A core assumption underpinning conceptualizations of temperament is that individual differences in rank-ordering on traits are relatively stable over time. However, it is also now recognized that temperament can show substantial rank-order change over time.

Correspondence: Daniel C. Kopala-Sibley, Ph.D., Department of Psychiatry, University of Calgary, Calgary, Alberta, 3280 Hospital Dr, NW, Daniel.kopalasibley@ucalgary.ca.

Understanding the degree and nature of this change in temperament over time is important for several reasons. Until recently, there was long-standing debate over the stability of temperament in terms of whether it was primarily environmentally or genetically influenced (See Caspi et al., 2005; Ferguson, 2010; Roberts & Delvecchio, 2000) although it is now generally recognized that both play a role and that there should be a moderate degree of both stability and change over time in temperament (Roberts & DelVecchio, 2000; Roberts & Mroczek, 2008). However, the degree of this stability varies depending on the developmental window examined, and rank-order stability from early childhood to early adolescence has been understudied. This issue is particularly important in children, who are sometimes assumed to show high rank-order levels of stability. For example, some may assume that a child prone to anger will likely always be more prone to anger than others their age.

The stability of temperament speaks to the heart of how of temperament is conceptualized and measured. For instance, it can inform knowledge about whether there are likely effects of developmental predictors of rank-order stability, and whether these likely have effects in specific developmental windows. For instance, many researchers have examined the effects of parenting on rank-order change in temperament over time (e.g., Kopala-Sibley et al., 2017; see Kiff, Lengua & Zalewski, 2011; Lipscomb, Leve, Harold, Neiderhiser, Shaw et al., 2011). However, if temperament shows substantial rank-order stability across childhood, this diminishes the likelihood of finding such effects during this period. Understanding the stability of temperament also has implications for identifying those at risk for negative outcomes early on. For example, given that negative emotionality is proximally related to depression and other psychological difficulties (see Klein et al., 2012), a child who is at risk based on their levels of negative emotionality may not remain at risk relative to others if their temperament is not stable over time. Finally, as Ferguson (2010) notes, this issue is not just relevant to psychological research and practice. There is a widespread belief that temperament is highly stable over childhood, which often leads to the assumption that any effort to change maladaptive temperament in a child is in vain.

While many studies have examined the stability of temperament in infancy and from early-to middle-childhood (e.g., Carranza et al., 2013; Durbin et al., 2007; Dyson et al., 2015; Komsi et al., 2008; Komsi et al., 2006; Rothbart, Derryberry, & Hershey 2000), few have examined the longer-term stability of temperament from early childhood to early adolescence. Such information is crucial to our understanding of the extent to which early temperament is stable over periods of more than a few years. Moreover, as the period from childhood through adolescence is characterized by critical changes in key psychological processes involved in individual differences (e.g., self-concept and identity, emotional reactivity, executive functioning) (Rothbart et al., 2000a; Sanson et al., 2004) and by evolving developmental pressures from the environment (e.g., Bakermans-Kranenberg et al., 2003), it is important to document how individual differences in traits manifest and change across this developmental span.

Another issue is that the vast majority of studies have relied exclusively on parent reports of child temperament (e.g., Lemery, Goldsmith, Klinnert, & Mrazek, 1999; Pedlow, Sanson, Prior, & Oberklaid, 1993; Roberts & Delvecchio, 2000; Rothbart, Derryberry, & Hershey, 2000). Fewer studies have used other assessment methods, such as laboratory and home

observations. As parent reports are only modestly associated with lab- and home-based observations of child temperament (Durbin et al., 2007) it is important to examine stability using multiple assessment approaches. Moreover, even less is known about the relationships of parent reports and observational measures of early childhood temperament with later child self-reports of traits (Mangelsdorf, Schoppe, & Buur, 2000).

Finally, rank-order stability in this literature is typically assessed via test-retest zero-order correlations, which may be attenuated due to measurement error at each time point while also inflated via shared method variance in terms of both informant and self-report methods (Ferguson, 2010; Roberts, Caspi, & Moffit, 2001). Studies examining the stability of latent temperament traits have been limited to toddlerhood or early to middle childhood (e.g., Dyson et al., 2015; Komsi et al., 2006, 2008; Neppl et al., 2010; Pedlow et al., 1993). Only one study of which we are aware has examined the rank-order stability of latent temperament traits based on both parent-report and lab-based observations (Majdandzic & Van den Boom, 2007), and that was limited to a seven-month assessment interval with 94 young children. To address these limitations and provide a comprehensive examination of the stability of temperament from early childhood to early adolescence, we examined the rank-order stability of lab-based observations of temperament as well as mother and father reports in a large sample of three-year olds with mother, father, and child reports of temperament in early adolescence.

Models of the structure of temperament

The content of temperament measures necessarily changes across developmental periods such that indicators in early childhood will be different from those in early adolescence (Neppl et al., 2010). As such, it is necessary to measure developmentally appropriate manifestations of the same traits at different ages. Theoretical models that specify core aspects of temperament are necessary to guide such work. The three-factor model proposed by Rothbart and colleagues (Rothbart, 1981; Putnam et al., 2001) is arguably the most widely-used contemporary model of childhood temperament (De Pauw & Mervielde, 2010; Zentner & Bates, 2008). It comprises the higher order factors of surgency, including high intensity pleasure, activity, impulsivity, and shyness (reversed); negative emotionality, including discomfort, anger/frustration, sadness, fear, and soothability (reversed); and effortful control (or constraint), including inhibitory control, attentional focusing, perceptual sensitivity, and low intensity pleasure (Rothbart et al., 2001). This model maps fairly closely onto taxonomies of personality and temperament derived from studies of adults (Caspi et al., 2005; Rothbart et al., 2000a), the best well-known of which are the five-factor and three-factor models (John et al., 2008, Markon et al., 2005, Tackett et al., 2008; Watson et al., 1994). Tellegen's (1985) influential three-factor model comprises positive emotionality, or a tendency towards positive affect and engagement with the environment; negative emotionality, or a propensity towards anxiety, anger, and fearfulness; and effortful control/ constraint, or, at the opposite pole, disinhibition, or a tendency toward impulsivity, risk-taking, and unconventional behavior. These higher-order trait dimensions encompass and explain many of the narrower trait dimensions from the different models of child and adult temperament (Caspi et al., 2005; De Pauw & Mervielde, 2010; Markon et al., 2005; Zentner & Bates, 2008).

Temperamental stability in younger children

Although it is somewhat unclear how rank-order stability of temperament in early childhood will extend to adolescence, shorter-term studies can inform expectations (Roberts & Delvecchio, 2000). A meta-analysis of longitudinal studies found average respective retest stabilities of .32, .52, and .45 across all traits in the periods from birth to three years, three years to 6 years, and 6 years to 12 years old (Roberts & Delvecchio, 2000). However, these estimates were not adjusted for measurement error, relied primarily on parent-reports of childhood temperament, and did not examine stability from early childhood to early adolescence.

More recent studies have reached similar conclusions about the rank-order stability of temperament from early to middle childhood. Several studies have examined the stability of temperament via latent variables and reported modest (~.20 - .35) to moderate (~.35 - .50) stabilities from early to middle childhood for negative emotionality, positive emotionality, and effortful control (e.g., Carranza et al., 2013; Durbin et al., 2007; Dyson et al., 2015; Komsi et al., 2008; Komsi et al., 2006; Neppl et al., 2010; Rothbart et al., 2000). Most of these studies, too, have similar limitations as those in Roberts & Delvecchio's (2000) meta-analysis in that they rely exclusively on one method such as parent-reports (Komsi et al., 2006; Komsi et al., 2008; Neppl et al., 2010; Rothbart et al., 2000) or lab reports (Dyson et al., 2015), do not adjust for measurement error by creating latent variables (Rothbart et al., 2000), and/or examine stability over relatively limited periods of development (Carranza et al., 2013; Durbin et al., 2007; Dyson et al., 2015; Komsi et al., 2006; Komsi et al., 2008; Rothbart et al., 2000).

We are aware of only two studies that have followed a sample from early childhood through early adolescence or beyond. Guerin and Gottfried (1994) found that parent reports of child temperament showed non-significant to moderate stability from age 2 to 12 years. Specifically, examining lower order facets of temperament, they reported test-retest correlations ranging from .00 for adaptability and .09 persistence to .23 (mood) to .30-.40 (activity, approach, intensity, and distractibility). However, they did not incorporate child reports or behavioural assessments or create latent variables to adjust stability estimates for measurement error. Similarly, in what is to our knowledge the longest-term study to examine continuity of early childhood temperament, Caspi & Silva (1995) used observers' ratings of temperament in 3-year old children completing cognitive and motor tasks to cluster children into groups, and found significant differences between subgroups on self-reports of positive and negative emotionality and constraint (versus disinhibition) at age 18 years. However, as noted by Caspi and Silva (1995) and Caspi et al. (1995), it is unclear whether these motor and cognitive tasks adequately elicited the range of behaviors of interest. Moreover, they did not use constructs derived from contemporary models of temperament, making it difficult to compare their findings to the current literature. For instance, their "inhibited" cluster of children showed a range of behaviors that are consistent with low positive emotionality, high negative emotionality, and high effortful control. At the age 18 follow-up, they relied solely on participant-reports, whereas parent-reports may still have provided useful information about their late adolescent children's temperament. Analyses in both studies also did not adjust for measurement error, and while the long-term follow-up in their study is a

significant strength, it is unclear how results would extend from early childhood to pre- rather than late-adolescence. Finally, neither of these studies integrated standardized lab-based assessments of early childhood temperament with parents' reports. Thus, knowledge of the rank-order stability of temperament from early childhood to early adolescence is surprisingly limited.

The current study builds upon previous work by integrating, at age 3, standardized lab-based assessments of temperament as well as widely-used questionnaires from both parents, both of which are grounded in well-established contemporary models of temperament. Further, we integrated child, mother, and father reports in early adolescence using a widely-used, well-validated measure of temperament that was designed to map on to the most widely-accepted three-factor models of child temperament.

Integrating multiple sources of information to study the stability of childhood temperament

Parents' reports are immensely valuable because they draw on observations over extensive periods of time and in a variety of contexts. However, they may be confounded by a variety of reporting biases (Christensen, Margolin, & Sullaway, 1992; Durbin & Wilson, 2012; Jensen, Traylor, Xenakis, & Davis, 1988; Webster-Stratton, 1988; Youngstrom, Izard, & Ackerman, 1999), and stability may be inflated by shared method variance and stability of parental perceptions rather than child behavior. Another approach to assessing temperament uses lab-based tasks designed to evoke the affects and behaviors characterising different temperament traits (Goldsmith et al., 1995). However, this approach also has potential limitations, such as concerns about ecological validity, although parents consider their child's responses during laboratory temperament assessments as highly typical of their behavior outside the laboratory (Lo et al., 2015). There may also be transient influences such as mood states as well as restrictions in the range of affect and behavior elicited in the child. As parent reports are only modestly associated with observational measures of child temperament, it is likely that both approaches provide unique perspectives (Durbin et al., 2007; Mangelsdorf et al., 2000).

In older youth and adults, temperament is often assessed via self-report. Self-reports can provide critical information about traits often unknown to other informants, such as parents and observers. Moreover, adolescent self-report instruments can map directly onto measures of adult personality and temperament, potentially providing a bridge between assessments of early-childhood and adult temperament.

We propose that the most informative and comprehensive approach to understanding the stability of temperament is to integrate lab-based observations with both mother and father reports in early childhood, and to examine their association with mother, father, and child reports in early adolescence. Each method provides important and informative but potentially incomplete information regarding children's temperament. This approach will also minimize shared method variance, thereby yielding more accurate estimates of the true rank-order stability of temperament over this span of development. Including both parents' reports of their child's temperament also mitigates parent-specific biases. Lab-based measures also provide an objective measure, whereas parents often have no comparison or norms by which to answer questionnaires. By measuring temperament via latent constructs

at each age, analyses also avoid attenuation of stability estimates due to measurement error. Finally, a latent variable modeling approach also permits the extraction of variance shared between each parent's report and lab-observations at age 3, and mother, father, and child reports at age 12, presumably providing a more valid measure of temperament.

There are, however, potential limitations to this approach. It is assumed in structural equation modeling (SEM) that the non-shared variance that is not included in the latent variable is measurement error, rather than variance in the true temperament construct. However, it is also possible that the non-shared variance provides valid and informative informant- or method-specific information. That is, while SEM is widely considered an appropriate way to assess the variance in an underlying construct that is imperfectly measured by several observed constructs, there may be valid and important variance that is lost by this approach. In addition, the present study examined only two time points and, due to the lengthy developmental span, was forced to use different measures at each occasion. Hence, it is limited to examining rank-order stability. Identical measures and three or more observations would be required for examining growth trajectories over time.

Using confirmatory factor analysis (CFA) to construct temperament and personality models with multiple latent factors can also be challenging (Hopwood & Donnellan, 2010). Exploratory factor analyses (EFA) of multi-factorial personality measures typically yield a good fit to the data as they allow cross-loadings of items and scales across multiple latent factors. However, because of these cross-loadings, CFA approaches to modeling multi-factorial personality measures typically show a poor fit to the data (Marsh et al., 2004; Marsh et al., 2010; Hopwood & Donnellan, 2010). More complex models that explicitly incorporate latent methods factors (e.g., Podsakoff, Mackenzie, Lee, & Podsakoff, 2003) often fail to identify or converge. As such, in the current study, we constructed latent variables in separate models for each of the Big Three temperament traits.

Measurement of temperament in early adolescence.

Given that the same measures of particular temperament traits cannot be used across widely spaced developmental periods, it is necessary ensure that the different measures used validly assess those constructs. In the current study, at the age 12 assessment, the Schedule for Nonadaptive and Adaptive Personality for Youth (SNAP-Y; Linde et al., 2013) was administered. The adult SNAP and SNAP-Y temperament scales were derived from the General Temperament Survey (Watson & Clark, 1992), which was based on Tellegen's (1985) three factor model of temperament, and contain the three factors of negative emotionality, positive emotionality, and disinhibition, with the latter mapping onto effortful control (Rothbart & Bates, 2006). Clark and Watson (2008) have shown that the Big Three factors meet all the criteria generally used to define temperament, in that they refer to patterns of affect and behavior that are relatively stable over time and are largely heritable but also influenced by the environment. Moreover, Rothbart and Bates (2006) and Shiner and Caspi (2003) note the substantial continuity between three-factor models of temperament in children, as examined in the current study, and three-factor models of personality and temperament in adults. These theorists argue that these three temperament dimensions are the foundation for later personality, and ultimately differentiate over

development in a broader array of personality traits, such as the Big Five (Ready & Clark, 2002; Rothbart & Bates, 2006).

Overview and hypotheses

The goal of this study was to examine the stability of childhood temperamental negative emotionality, positive emotionality, and effortful control/constraint over a 9-year period from age 3 years to age 12. First, we examined zero-order correlations of each trait across different measurement methods and informants. We expected to find the strongest convergence when assessed both within-informant and within-construct. We also expected to observe significant but weaker convergence when the same construct was assessed via parent report but across mothers versus fathers. We expected the lowest, but still significant, convergence between lab-based measures at age 3 and parent reports at age 12. Given the lack of prior research on this topic, we had no *a priori* hypotheses for whether lab- versus parent-reports at age 3 would best predict age 12 child reports. Analyses also examined whether age 3 constructs correlated more strongly with their corresponding age 12 counterpart than with non-matching temperament traits (i.e., discriminant validity).

Second, we constructed latent temperament constructs at age 3 based on mother and father reports as well as lab-based observations. We also constructed latent age 12 temperament constructs based on mother, father, and child reports. We then examined rank-order stability between these latent constructs. We expected these stability coefficients to be moderate in strength, but higher than the bivariate test-retest correlations. By incorporating multiple informants and multiple methods and following a large cohort of 3-year old children over a 9 year period, the current study represents the most comprehensive examination of the stability of early childhood temperament of which we are aware.

Methods

Open materials and open data statement

Syntax, AMOS input files, data files, and scripts for all analyses have been made available.

Participants

Participants were 559 three-year-old children and their parents living in Long Island, New York, who were recruited as a part of a longitudinal study of children's temperament (see Olino, Klein, Dyson, Rose, & Durbin, 2010 for details). The mean age of the children at baseline was 43.5 months (SD = 2.8).

In 2004–2007, participants were recruited through a commercial mailing list and screened by phone. Eligible children had no significant medical problems or developmental disabilities, and had at least one English-speaking biological parent who could participate. Most children were European American and non-Hispanic (86.9%), lived with both biological parents (94.6%), and came from middle-class families, as measured by the Four Factor Index of Social Status (M = 45.33, SD = 10.99; Hollingshead, 1975). Families in this study had an average annual household income of approximately $100,000, which is highly

comparable to the average household income in the geographic area from which this sample was drawn.

The effective sample size was based on the number of children who completed the Lab-TAB at baseline. Of the 559 families whose children completed the Lab-TAB at baseline, 41 mothers and 158 fathers did not complete parent-report temperament measures. Experimenters also rated participants' traits across the entire lab visit ("Experimenter-Impressions"); these scores were unavailable for 22 children. Participants with missing data at baseline did not differ from those with complete data for mothers, fathers, and lab-based observations in terms of sex, socioeconomic status (SES), living in a two-parent home, race, ethnicity, or any measure of temperament at age 3 (all ps > .05). Participants were assessed again at age 12 (2014–2016, M age = 151.95 months, SD = 5.53). Of the 559 original participants, 423 children, 426 mothers, and 359 fathers completed questionnaires at the age 12 follow-up. Participants with missing data at the age 12 assessment did not differ from those with complete data at both time points on any variable assessed in this study (all ps > .05). Little's MCAR test also confirmed that missingness was not significantly related to any variable in our study: χ^2 (436) = 482.103, p = .06. Data can thus be viewed as missing at random for analyses. Full Information Maximum Likelihood (FIML) procedures in AMOS 22.0 were used to estimate the means and intercepts in the presence of missing data. This approach is generally acknowledged to be preferable to listwise deletion or mean imputation, which are more likely to yield biased estimates (e.g., Schafer & Graham, 2002). Our effective sample size in latent models was therefore 559, although sample sizes for zero-order correlations varied.

Procedure

At age 3, children participated in the Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith et al., 1995). Mothers and fathers completed the Child Behavior Questionnaire (CBQ; Rothbart et al., 2001). At age 12, mothers, fathers, and children completed the adaptive temperament subscales of the Schedule for Non-adaptive and Adaptive Personality for Youth (SNAP-Y; Linde et al., 2013). See Table 3 for a summary of the variables, measures, and constructs assessed at each time point.

Materials

Lab-based observations of temperament.—At age 3, child negative emotionality, positive emotionality, and Impulsivity were coded by research assistants based on videotapes of the Lab-TAB as well as experimenter-impression ratings. Before coding independently, coders had to achieve 80% or higher agreement with an expert coder on all codes within an episode. Twelve age-appropriate tasks were used (Dyson, Olino, Durbin, Goldsmith, Bufferd et al., 2015; Olino et al., 2010). Most were adapted from tasks used in the developmental literature and were designed to elicit a range of temperament-relevant emotions and behaviours from the child (Gagne, Van Hulle, Aksan, Essex & Goldsmith, 2011; Goldsmith et al., 1995). For a full description of all episodes at each assessment wave, see Appendix A.

During each task, each instance of children's bodily, vocal, and facial expressions of positive affect, sadness, anger and fear were rated on a 3-point scale (i.e. low, moderate, high

intensity). Intensity ratings were then summed for each channel (i.e. facial, bodily, and vocal) within each episode. Following this, the intensity ratings were averaged within each channel across all episodes for each of the affective traits. Standardized scores for the three channels were then averaged to comprise the composite score for each affect. Interest/ engagement was rated on a 4-point Likert scale based on each full episode and reflected the child's tendency to display interest in stimuli, ask questions or make comments, and their general level of engagement. Impulsivity/disinhibition was also rated on a 4-point Likert scale based on the entire episode, and reflected the child's tendency to act or respond without reflection or hesitation, as indicated by quick changes in behavior or shifts in attention. Impulsivity was reversed so that scores were in the same direction as effortful control subscales of the CBQ. Standardized ratings of sadness, anger, and fear were combined to create a total negative emotionality scale, and standardized ratings of positive affect and interest/engagement were composited to create a positive emotionality scale. Coefficient alphas for negative and positive emotionality were both .82, and the intraclass correlation coefficient (ICC) for interrater reliability (n = 35) for negative and positive emotionality was .74 and .89, respectively. Alpha for impulsivity was .70, with an ICC of .75. The Lab-TAB shows moderate test-retest stability and construct validity in terms of associations with independent ratings by experimenters, unstructured home observations, and diagnostic interview assessments of child psychopathology (Dougherty et al., 2011; Durbin et al., 2007; Gagne et al., 2011).

The individual conducting the laboratory visit with the participant completed a set of global ratings about the child (Experimenter Impressions). These were adapted from the post-Lab-TAB rating scale developed by Gagne et al. (2011). The experimenter was with the child from the moment the family arrived on campus until they walked the participant and parent back to the parking garage after the assessment, including during the unstructured play breaks between each of the 12 Lab-TAB episodes. Thus, experimenter ratings are both more global and based on a larger sample of child behavior. Experimenters rated the participant on 24 scales, 15 of which were used here, and from which we derived overall measures of negative emotionality, positive emotionality, and effortful control/constraint. Each variable was rated on a single 5-point Likert scale (1=rarely, 2=subtle or ambiguous signs, 3=mild, 4=moderate, 5=extreme). We averaged ratings of relevant items to create temperament scales. Negative emotionality was comprised of overall negative affect, fearfulness, frustration with tasks, anger or irritability, and sadness (Cronbach alpha = .79). Positive emotionality was comprised of overall positive affect, interest in test materials and stimuli, enthusiasm toward tasks, initiative, and anticipatory positive affect (alpha = .89). Effortful control/constraint was comprised of attention to tasks, persistence in completing tasks, and impulsiveness (alpha = .75). Interrater reliabilities are not available as there was only one experimenter per Lab-TAB assessment.

Parent-report of child temperament at age 3.: Mothers and fathers completed the CBQ (Rothbart et al., 2001) at baseline. The CBQ is currently the most widely-used parent-report measure of early childhood temperament (Putnam et al., 2006; Rothbart et al., 2001). It is a 195-item questionnaire, with each item rated on a 7-point Likert scale ranging from "extremely untrue of your child" to "extremely true of your child." Responses are averaged

to create a total score. The present study used the higher-order effortful control (EC) and negative affectivity (NA) scales. However, surgency, the third higher order factor on the CBQ, has broader content than the construct of positive emotionality. Therefore, we used the CBQ lower order facets of approach/anticipation and smiling/laughter as indicators of positive emotionality (PE) as these are the two CBQ subscales which most closely correspond to this trait. In prior research these scales have shown good internal consistencies and test-retest reliability, and correlations with expected outcomes such as social behavior (Rothbart et al., 2001), although it should be noted that recent CFA-based research has not supported the proposed structural properties of the CBQ (Kotelnikova et al., 2016). In the current sample, alphas for each subscale used in this report ranged from .63 (sadness) to .79 (anger/frustration).

Age 12 Temperament.: Age 12 temperament was assessed via mother-, father-, and self-reports about the child on the adolescent version of the SNAP (Linde et al., 2013). The SNAP-Y is a 390-item factor-analytically derived measure with true/false response options. It contains 3 temperament scales, 15 lower-order personality trait dimensions across the continuum from normal to abnormal personality functioning, and 6 validity scales. We administered only the three temperament scales: negative temperament (referred to here as negative emotionality), positive temperament (referred to here as positive emotionality), and disinhibition. Negative emotionality (28 items) assesses tendencies towards anger, sadness, fear, and distress (e.g., "I rarely get so angry that I lose control," and "I often feel nervous and tense"). Positive emotionality (28 items) assesses tendencies towards positive affect, activity, enjoyment, and pleasure (e.g., "I lead an active life," and "I get excited when I think about the future"). Disinhibition (35 items) assesses the extent to which behaviors are planned, thought through, and not impulsive (e.g., "I rarely, if ever, do anything reckless," and "I never buy things on a whim or impulse"). Thus, these constructs correspond closely with the temperament constructs assessed at age 3 by the Lab-TAB and CBQ.

The SNAP and SNAP-Y show good convergent validity with a range of normal and abnormal personality traits and forms of psychopathology (e.g., Clark, McEwen, Collard, & Hickok, 1993; Linde et al., 2013; Melley et al., 2002; Watson, 2000; Watson & Clark, 1992). In the current study, alphas for child, mother, and father reports of negative emotionality were .90, .88, and .89, respectively. For positive emotionality, child, mother, and father reports had alphas of .80, .85, and .86, respectively. For disinhibition, alphas for child, mother, and father reports were .80, .84, and .83, respectively. Correlations with other individual differences measures in this study support the convergent and discriminant validity of the SNAP-Y, and are provided in Supplementary Table S6.

The child and the parent attending the lab-visit completed the SNAP-Y on a computer during that lab visit, while the other parent completed it electronically at home. There were no time restrictions.

Data Analyses—Analyses initially consisted of zero-order correlations between Age 3 and 12 temperament variables which were corrected for attenuation due to unreliability (see Muchinsky, 1996), as measured by Cronbach alpha values of internal consistency. As noted by Ferguson (2010) in his meta-analysis, zero-order correlations as measures of the stability

of temperament are attenuated by 19%–26% percent because of unreliability. The formula for correcting for attenuation is $r_{xy}/ (r_{xx}*r_{yy})$ where r_{xy} is the raw correlation between x and y, r_{xx} is the Cronbach alpha of x, and r_{yy} is the Cronbach alpha of y. Uncorrected correlations and their two-tailed p-values are shown in, with the corrected correlations in. Adjusting a correlation for unreliability does not alter its p-value. These correlations examined the convergence and divergence of measures of temperament across methods, informants, and variables from age 3 to 12. Where an age 3 trait correlated with multiple informants' reports at age 12 on the same trait, or where multiple informants' reports of a specific age 3 traits correlated with a particular age 12 score, correlations were z-scored and compared in order to test the significance of the difference in the strength of the associations. Significant differences between correlations are shown in, while results of all comparisons between correlations are shown in Supplementary Tables S3-S5.

Structural Equation Modeling was carried out using AMOS 22.0. We examined the rank-order stability between temperament at age 3 measured via a latent variable indicated by mother and father CBQ reports, and Lab-TAB and experimenter impression scores, and at age 12 via a latent variable indicated by mother, father, and child reports on the SNAP-Y. Thus, age 3 and 12 temperament were modeled simultaneously. At age 3, latent negative emotionality was indicated by mother and father reports of CBQ negative affectivity as well as negative emotionality from the Lab-TAB and Experimenter Impressions. Latent positive emotionality was indicated by mother and father reports of smiling-laughter and approach-anticipation on the CBQ, and Lab-TAB and Experimenter Impressions positive emotionality. Effortful control was indicated by mother and father reports of effortful control on the CBQ with LabTAB impulsivity (reversed) and Experimenter Impressions effortful control/ constraint. At age 12, latent variables of negative emotionality, positive emotionality, and disinhibition were each comprised of mother, father, and child reports on the relevant manifest indicator. Given that ratings from Experimenter Impressions and the Lab-TAB were based on overlapping samples of behavior, error terms on Lab-TAB and Tester-Impressions variables were covaried within each construct *a priori* in each model.

Ideally, the interrelationships between all three latent factors at both age 3 and 12 should be examined in the same model. However, consistent with prior findings that CFA approaches to modeling the structure of personality or temperament typically show a poor fit to the data (Marsh et al., 2004; Marsh et al., 2010; Hopwood & Donnellan, 2010), a three factor model of temperament at age 3 including covariances on error terms within each construct of lab-based variables a showed a poor fit to the data, $\chi 2 (71) = 694.70$, $p < .001$, CFI = .70, RMSEA = .13, 90% CI [.12, .13].

We also attempted to model all three age three temperament traits in one model following Podsakoff's proposed multi-trait multi-method CFA (see in Podsakoff et al. [2003]). Latent variables were indicated by separate mother, father, and lab-based scores, thereby portioning variance attributable to the informant or method into separate latent factors, while retaining separate latent factors for negative emotionality, positive emotionality, and effortful control (Supplementary Figure S1). We also computed a similar model but with only two latent methods factors, one for lab-based variables and another for all parent-reports (Supplementary Figure S2). However, consistent with Podsakoff et al.'s (2003) observation

that the most frequent problem with these models is that they do not identify, neither model identified. Others have also noted frequent convergence problems with these models (see Kenny & Kashy, 1992).

We also computed Podsakoff's correlated uniqueness model with all three traits at age 3 (in Podsakoff et al., 2003). This model accounts for methods effects by allowing error terms of specific measurement methods to be correlated. Thus, covariances were included between all parent-report variable error terms, and between all lab-measured variable error terms (see Supplementary Figure S2). Again consistent with Podsakoff et al.'s (2003) point that these models often have identification problems, this age 3 model failed to identify. A three factor model of SNAP-Y scales at age 12 similarly showed a poor fit to the data, $\chi2$ (24) = 176.23, p < .001, CFI = .84, RMSEA = .11, 95% CI [.09, .12]. Given that we were unable to model all three latent factors in one model at age 3, we did not compute similar models with age 12 variables. Therefore, we examined the stability of each trait in separate models.

In order to examine heterotypic continuity, we imputed factor score estimates from the model for each of the latent variables described above based on the factor loadings from our models that contained only one age 3 and one age 12 latent variable. That is, negative emotionality at ages 3 and 12, positive emotionality at ages 3 and 12, and effortful control at ages 3 and 12 were imputed separately, and each was used as a manifest (observed) variable in our models examining heterotypic continuity. AMOS uses regression imputation with FIML for estimating missing data. This is identical to applying the factor score weights to each individual's scores on the observed indicators. Estabrooke and Neal (2013) note that when data is Missing at Random (MAR), FIML more accurately estimates individual factor scores with missing data compared to sum scores, mean imputation, or regression estimators that do not use FIML. We then tested a path model in which each Age 12 latent temperament variable (i.e., negative and positive emotionality and disinhibition) was simultaneously regressed on all three Age 3 latent temperament variables (i.e., negative emotionality and positive emotionality and effortful control). Thus, this analysis examined whether each age 3 trait predicted each age 12 trait, adjusting for that trait's respective age 3 levels.

We are not aware of any studies which have used multiple methods and informants and created factor scores estimates of the latent variables based on these observed indicators in separate models. However, it is routine for constructs to be created based on averages of either parent reports or behavioral episodes, and then use those to examine heterotypic stabilities of over time (e.g., Majdandzic & Van den Boom, 2007; Pesonen et al., 2003; Putnam et al., 2008). Komsi et al. (2006) examined heterotypic continuity of latent temperament constructs, but these constructs did not include observational measures as indicators, and so they were able to fit multi-factorial temperament models. As such, we are not aware of any studies which have taken an approach directly comparable to ours.

As measures of goodness of fit, we present chi-square, comparative fit index (CFI), and root-mean-square error of approximation (RMSEA) as well as the 90% confidence interval around the RMSEA. Generally, CFI values greater than .90 (Hoyle & Panter, 1995), and an RMSEA less than .08 (Kline, 1998) indicate acceptable fit. Given that we expect a medium effect size in our latent stability models both in terms of loadings of observed indicators on

latent variables as well as the relationships between age 3 and 12 temperament traits, power analyses revealed that with two latent variables and 9 indicators, which is the upper number of indicators in our models, a minimum sample of 90 would be required (Westland, 2010; Soper, 2017). Regression weights in our latent models are fully standardized and therefore represent effect sizes.

Results

Zero-order convergent and discriminant correlations

Descriptive statistics and within-time correlations are available in supplementary Tables S1 and S2. Zero-order correlations between age 3 lab-based and parent-reported temperament traits and age 12 mother, father, and child-reported traits are shown in, and associations unadjusted for unreliability are presented in.

Negative emotionality correlations.—Consistent with hypotheses, correlations were generally strongest between parent reports of negative emotionality at age 3 and 12 compared to correlations of age 3 lab-based negative emotionality with age 12 negative emotionality, with some evidence of stronger correlations within compared to across informants (and S3). Also consistent with expectations, most correlations were modest in strength. Compared to parent reports of negative emotionality at age 3, Lab-TAB and experimenter impressions negative emotionality converged as well or better with child-reported negative emotionality at age 12. Lab-TAB negative emotionality also correlated negatively with mother-reported age 12 positive emotionality, while mother-reported age 3 negative emotionality correlated positively with mother-rated age 12 disinhibition.

Positive emotionality correlations: Also consistent with expectations of stronger within-than across-informant correlations, parent-reports of positive emotionality at age 3 generally only converged with their own reports of positive emotionality at age 12 (and S4). These associations were also modest in strength. In contrast, Lab-Tab positive emotionality converged with both mother and father reports of positive emotionality at age 12. Interestingly, Lab-TAB positive emotionality showed somewhat better discriminant validity in that it was related only to age 12 positive emotionality variables, whereas parent reports of age 3 positive emotionality correlated with several measures of negative emotionality and disinhibition at age 12.

EC correlations.: Both Lab-based measures and parent-reports of effortful control at age 3 showed good convergence with all informants' reports of age 12 effortful control (and S5). As expected, several of these correlations were stronger within-informants than across informants or measures, although both the lab-based and parent-report measures converged highly with age 12 child-reports of effortful control. However, lab-based measures of effortful showed better discriminant validity in that they uniquely correlated with age 12 effortful control. In contrast, parent-reported age 3 effortful control showed multiple associations with measures of age 12 negative and positive emotionality.

Stability of latent constructs

Negative emotionality.: Our model of latent age 3 negative emotionality predicting latent age 12 negative emotionality (Figure 1, top panel) showed a good fit to the data: χ^2 (12) = 33.97, p = .001, CFI = .96, RMSEA = .054, 90% CI [.034, .079]. All loadings of indicators on age 3 and age 12 negative emotionality were at least .20 and were significant at $p < .001$, although CBQ variables showed higher loadings than Lab-TAB or Experimenter Impressions negative emotionality. The standardized stability of age 3 to age 12 negative emotionality was .41 (p = .003), indicating moderate stability.

Positive emotionality.: Our model of latent age 3 positive emotionality predicting latent age 12 positive emotionality (Figure 1, middle panel), had an adequate fit: χ^2 (25) = 89.16, $p < .$001, CFI = .91, RMSEA = .068, 90% CI [.053, .083]. All loadings of indicators on age 3 and age 12 positive emotionality exceeded .26 and were significant at or below p = .001. The standardized stability of age 3 to age 12 positive emotionality was .30 (p = .001), indicating modest stability.

EC. Our model of latent age 3 effortful control predicting latent age 12 disinhibition (Figure 1, bottom panel) showed a good fit to the data, χ^2 (12) = 36.21, $p < .001$, CFI = .98, RMSEA = .06, 90% CI [.036, .083]. All loadings of indicators on age 3 effortful control and age 12 disinhibition exceeded .33 and were significant at or below p = .001. Higher scores on Age 3 effortful control indicate higher levels of effortful control. The standardized stability of age 3 effortful control to age 12 disinhibition was −.53 ($p < .001$), indicating moderate to high stability. The stability of disinhibition was significantly greater than that of both positive emotionality (z = 2.41, p = .02) and negative emotionality (z = 2.82, p = .005). The stability of positive and negative emotionality did not significantly differ ($p > .30$).

Heterotypic continuity.: Using the imputed age 3 and 12 temperament scores from our latent models, we examined whether each age 3 temperament trait predicted each discriminant age 12 temperament trait after adjusting for each age 12 trait's respective level at age 3 (e.g., age 3 negative emotionality predicting age 12 disinhibition adjusting for age 3 effortful control). Given that our model examining heterotypic continuity was fully saturated, fit indices are perfect and are therefore not reported. In this model (Figure 2), in addition to significant homotypic stability coefficients, greater age 3 effortful control predicted increases in positive emotionality at age 12 and decreases in negative emotionality at age 12. Including heterotypic paths substantially attenuated the homotypic paths.

Ancillary analyses.: In order to ensure robustness of our results and stability estimates, latent stability models as well as the heterotypic continuity model were recomputed after deleting any participants who had missing data on any variable at either time point. This resulted in a sample of 260 with complete data on all variables. Results were quite similar to those from the full sample with missing data estimated. For positive emotionality, the model showed an adequate fit, χ^2 (29) = 69.36, $p < .001$, CFI = .90, RMSEA = .08, 90% CI [.060, .106], with a standardized stability estimate of .41. For negative emotionality, the model showed an adequate fit, χ^2 (12) = 27.99, p = .006, CFI = .95, RMSEA = .07, 90% CI [.037, .107], with a standardized stability estimate of .42. For effortful control, the model showed

an adequate fit, χ^2 (12) = 28.83, p = .004, CFI = .98, RMSEA = .07, 90% CI [.039, .110], with a standardized stability estimate of −.57.

Results using listwise deletion for models examining heterotypic continuity were also similar to those originally reported using all available data. Age 3 effortful control predicted lower NT at age 12 (β = −.19, p < .001) and greater PT at age 12 (β = .15, p = .004). There was also a newly significant effect of greater NE at age 3 predicting decreases in age 12 PT (β = −.14, p = .006), over and above age 3 PE. However, given that estimating missing data via FIML has been shown to reliably produce similar regression estimates as complete data sets (Schafer & Graham, 2002), we do not interpret this finding.

Discussion

This study examined the stability of temperament over almost a decade ranging from early childhood through early adolescence in a large sample which integrated both mother and father reports as well as lab-based observations of early childhood temperament, and mother-, father-, and child-reports of temperament in early adolescence. At the zero-order level, mother- and father-report measures generally showed stronger convergence within than across informants, and stronger associations than did lab-based observations with age 12 parent-reports. However, lab-based observations at age 3 generally showed greater specificity than parent-reports in their associations with age 12 temperament, in that parent-reports of early childhood temperament often correlated with non-corresponding traits at age 12. Stability estimates derived from latent models suggested at least moderate stability in temperament from age 3 to 12. Finally, heterotypic analyses showed continuity from higher effortful control at age 3 to lower negative temperament and higher positive temperament at age 12. This suggests that effortful control, which is believed to facilitate regulation of emotional reactivity, influences the trajectories of negative and positive emotionality, both of which underpin children's emotional reactivity.

Zero-order stability and specificity of early- to late-childhood temperament

Mothers' and fathers' reports of child negative emotionality, positive emotionality, and effortful control at age 3 typically showed stronger convergence with their own report of the same trait at age 12 relative to the other parent's report. This is consistent with Neppl et al. (2010) who found stronger test-retest correlations within informants (mothers versus fathers) than across. Overall, results are consistent with the possibility that shared method variance, both in terms of method (questionnaires) and informant, increases estimates of stability. This is significant because most studies of the stability of temperament use the same measures and informants at each assessment.

Associations between age 3 parent-reported negative and positive emotionality with their respective age 12 child-reported counterparts showed some unexpected patterns. Father, but not mother, reports of negative emotionality at age 3 correlated with child-reported negative emotionality at age 12. Across mother and father reports of smiling-laughter and approach-anticipation, only father-reported smiling-laughter correlated with child-reports of age 12 positive emotionality. Many would likely assume that a mother might have more detailed and accurate information regarding their young child's negative and positive emotionality.

However, if child reports at age 12 are taken as an important marker of their age 12 temperament, this result suggests that fathers' reports in early childhood provide better information in terms of their correlation with child reports in early adolescence.

The possibility that fathers' have more detailed or accurate information compared to mothers regarding their young child's temperament, at least in terms of their convergence with children's reports in early adolescence, requires replication in future research. If verified, this may suggest that mothers should not be relied upon as the sole source of information regarding their child's temperament. These results may also suggest that fathers' reports of temperament show greater utility compared to mothers' reports in terms of their convergent validity with children's later reports. The reasons that fathers' reports showed greater convergent validity than mothers' with children's later self-reports cannot be gleaned from the current study. We speculate that fathers are more observant of, and attentive to, their young children's emotional expressions, which would be contrary to the oft-expressed position that mothers provide more accurate reports of child behavior. Alternatively, fathers may be more likely to notice or attend to more extremes in children's emotional behaviour, which could lead to better discrimination between more and less emotional children. If so, this could be driving the greater convergence between early childhood father reports and early-adolescent children's reports. Regardless, consistent with prior research highlighting that different informants provide unique sources of information (e.g., Achenbach et al., 1987), these results underscore the importance of incorporating fathers' reports in the comprehensive measurement children's temperament, whether for research or practical/ clinical purposes.

Mother- and father-reported age 3 effortful control correlated with age 12 child reports of disinhibition, although more weakly than with parents' reports of disinhibition at age 12. Experimenter Impressions and Lab-TAB positive emotionality as well as Experimenter Impressions negative emotionality were not significantly associated with child reports of their corresponding constructs at age 12, although Lab-TAB negative emotionality and both lab-based observations of effortful control were. These results suggest that lab-based measures, in particular those measuring effortful control and to some extent negative emotionality, tap the early antecedents of child-reports of temperament in early adolescence.

It is possible that the behaviours and affects associated with negative emotionality (e.g., sadness, anxiety) as well as effortful control (e.g., actively restraining an impulsive behaviour) are experienced internally to a greater degree than they are expressed externally by the child. Lab-based tasks may elicit anxiety (e.g., a stranger approaching), sadness (e.g., not receiving an expected gift), or effortful control (e.g., being forced to wait while building a tower) that may be indicative of these specific behaviors in naturalistic settings, and thus better converge with child-reports early adolescence. As such, results further suggest the importance of including lab-based measures of child temperament in addition to parental reports.

Parent-rated age 3 temperament constructs often showed poor discriminant validity with age 12 temperament. This is consistent with some prior research that has used parent reports (e.g., Neppl et al., 2010). In the current study, father-reported age 3 negative emotionality

showed the best convergent and discriminant validity, in that it correlated with mother, father, and child reports of negative emotionality at age 12 but not with any other age 12 variable. Mother reports of age 3 negative emotionality, however, correlated not only with mother and father age 12 negative emotionality, but also with mother reports of disinhibition at age 12. While parent-reported CBQ smiling-laughter and approach-anticipation generally showed some specificity with age 12 positive emotionality, mother reports of smiling laughter also correlated with mother-reported disinhibition at age 12. Parental reports of effortful control showed the poorest discriminant validity, in that they both correlated with parents' reports of negative emotionality and positive emotionality at age 12, in addition to correlating with disinhibition.

Lab-TAB and Experimenter Impressions temperament variables, on the whole, showed good discriminant validity. Lab-TAB positive emotionality was uniquely related to father and mother reports of age 12 positive emotionality, but was not significantly related to negative emotionality or disinhibition as rated by any informant. Similarly, Experimenter Impressions positive emotionality was uniquely related to age 12 mother-rated positive emotionality, but was unrelated to any other age 12 variable. Lab-TAB impulsivity and Experimenter Impressions effortful control were uniquely related to mother, father, and child reports of disinhibition, and were unrelated to any informant's report of positive or negative emotionality. Finally, Lab-TAB negative emotionality predicted child-reported age 12 negative emotionality, and Experimenter Impressions negative emotionality was specifically related to mother-reported age 12 negative emotionality. The one exception to the pattern of high discriminant validity for lab-based variables was that Lab-TAB negative emotionality at age 3 predicted lower levels of mother-rated positive emotionality at age 12.

Taken together, lab-based measures may show better discriminant validity in terms of their associations with age 12 temperament relative to parent reports of temperament in early childhood. One possibility is that parents may have somewhat misinterpreted the intent of the questions or their child's behaviour in everyday life. Similarly, the divergent correlations observed for parent reports may reflect shared method variance. Finally, we cannot rule out that what may appear to be a lack of discriminant validity is really greater sensitivity to true heterotypic continuity relative to lab-based measures. However, this would require systematic evidence across a variety of measures that particular traits predict other traits later in development. Regardless, these results further highlight the utility of incorporating lab-based measurements into the assessment of temperament in young children.

Stability of latent temperament constructs

Results from SEMs indicated that lab-based observations and mother and father reports of temperament in early childhood were viable manifest indicators of latent temperament constructs, although Lab-TAB negative emotionality and impulsivity and Experimenter Impressions negative emotionality and effortful control contributed relatively lower amounts of variance to their respective latent factors than the CBQ variables. Results suggest that assessing temperament via both parents as well as lab-based observations may yield a more comprehensive and nuanced measure than using only one of these methods.

Across all three temperament traits, SEMs indicated that temperament is moderately stable from age 3 through 12 years, although effortful control showed somewhat greater stability than did negative or positive emotionality. Our results are consistent with Caspi and Silva (1995) who reported significant associations between temperament at age 3 and personality at age 18, as well as Guerin and Gottfried (1994) who reported moderate stability of maternal reports of children's mood ($r = .23$), distractibility ($r = .37$), and approach ($r = .42$) from age 3 to 12.

Our stability estimates, based on latent variable modeling over a period of almost a decade were comparable to or higher than those reported by Roberts and Delvecchio (2000), despite it being reasonable to expect higher stability estimates over shorter time periods. That suggests that if one uses multiple sources of information to reduce measurement error, there is a moderate to substantial degree of continuity in temperament from early childhood to early adolescence, even though this period is believed to be characterized by greater flux in individual traits than subsequent developmental periods (Posner & Rothbart, 2007; Rothbart & Bates, 2006). Interestingly, latent stability coefficients were higher than the highest within-informant correlations. This suggests that random error may have reduced the zero-order correlations more than informant/method variance inflated it. We are not aware of studies that have attempted to quantify the relative effects of informant/method variance versus random error variance on the associations between variables.

Taken together, these results provide strong evidence for homotypic continuity as well as moderate rank-order stability in temperament. These findings speak to the fundamental nature of temperament in children. Temperament is arguably at the core of human psychological identity; our results suggest that this identity begins to solidify over the course of childhood while also remaining malleable. These findings support the *compromise perspective* of temperamental stability and suggest that temperament should be conceptualized as only moderate stable, at least across childhood. This is consistent with the possibility that there are important environmental influences, such as parenting, culture (Ferguson, 2010), and non-shared environmental factors (see Saudino, 2005) that affect the development of temperament. At the same time, the results suggest that early child temperament, at least when measured via multiple methods and informants, provides a marker with modest predictive utility of future temperament styles up to 9–10 years later.

Finally, the results from our path model also showed some evidence of heterotypic continuity. Specifically, greater effortful control predicted lower negative emotionality and higher positive emotionality at age 12. This is consistent with some previous research that found evidence of heterotypic continuity of temperament in early childhood. For instance, Putnam et al. (2008) infants high in surgency showed greater effortful control in toddlerhood, high levels of surgency in toddlerhood predicted lower effortful control in preschool, and toddler negative affect predicted lower effortful control in preschool (Putnam et al., 2008). Evidence also suggests that effortful control predicts lower levels of negative emotionality in later childhood (Kochanska & Knaack, 2003). Similarly, Dyson et al. (2015) also found that age 3 constraint, similar to effortful control, predicted lower age 6 fearfulness. The reasons that early childhood effortful control influences later negative and positive emotionality in early adolescence cannot be gleaned from the current study.

However, this is consistent with prior evidence that effortful control is associated with more effective regulation of emotional reactivity, and more generally with positive outcomes in later life (e.g., Eisenberg et al., 2009; Moffitt et al., 2011). Results suggest the need to search for potential mechanisms underlying homotypic and heterotypic continuity of temperament, such as specific genetic and environmental influences, gene-environment interactions, and gene-environment correlations (Caspi & Roberts, 2001; Plomin et al., 1999).

Our results also highlight the importance of considering multiple sources of information regarding children's temperament in addition to parent reports. More generally, the results suggest that researchers should carefully consider the methods they use to assess temperament, as well as how they interpret their results based on those methods.

Limitations and future directions.

This study had several notable strengths. It comprised a large sample of children who were followed for almost a decade, and used fine-grained, standardized, lab-based measures of temperament along with both mother and father-reports in early childhood, and mother-, father-, and child-reports of temperament in early adolescence. However, some limitations should be noted. First, scores on the Lab-TAB and Experimenter Impressions variables may be influenced by situation-specific behaviors or affects. However, prior research suggests that the stability of lab-based observations of negative and positive emotionality is almost comparable to parent reports (Durbin et al., 2007).

Second, we were unable to assess the interrater reliability of the Experimenter Impressions measure. However, the integration of these variables with Lab-TAB ratings and parent-reports in our latent models attenuates this concerns. Experimenter Impressions ratings were also derived from observations that overlapped with the Lab-TAB, which showed good interrater reliability and were moderately-highly correlated with the Lab-TAB (see Supplementary Table 1).

Third, we also could not use lab-based measures at age 12, given that there are no validated observational measures of temperament in older youth. This raises the concern of whether stability estimates, even those derived from latent variables, may underestimate stability.

Fourth, we only examined rank-order stability. Future research should also examine mean level stability as well as individual trajectories over time in temperament. However, this will be challenging, as these questions require the same measures to be administered at each time point, which is difficult given considerations of developmental appropriateness. The lack of similarity of indicators in our latent models also precludes examining measurement invariance over time. This is a fundamental challenge to the field, as measures need to be developmentally appropriate while also still measuring the appropriate construct. This typically precludes using identical measures when developmental samples are followed across different development periods, as in the current study.

Fifth, although not a limitation per se, we did not examine predictors of change in temperament over time or moderators of the stability of temperament. While we hope to examine these issues in future research, prior to examining influences on stability it is

important to first understand normative stability. Sixth, future research would also benefit from examining stability of lower order temperament traits.

Seventh, the participants in our sample were predominantly White/European American and middle class and recruited through a commercial mailing list. Although the sample was demographically representative of the population in our geographic region (Suffolk County, Long Island, New York), this may constrain the generalizability of our findings.

Finally, were unable to test heterotypic continuity in our full latent models, and instead had to circumvent this by imputing latent scores and using them as observed variables. It is unclear whether results would change using a multi-factorial confirmatory model. For instance, by estimating factors scores based on separate models for each construct, this may bias results in favor of finding homotypic paths as estimated scores were based on homotypic latent models that examined only one trait each.

Conclusion

Understanding the extent to which childhood traits are stable has important implications for our knowledge of normative development of individual differences and core aspects of how we view ourselves and others. Few studies have examined the stability of early childhood temperament through to early adolescence and integrated lab-based observations of temperament with mother and father reports in early childhood and mother, father, and child reports of temperament in adolescence. Zero-order correlations suggested that associations over time are influenced by a variety of considerations, in particular the assessment method and informant at baseline and follow-up. Latent models suggest that stability estimates increase after removing measurement error and showed moderate stability (bs = .30 - .55) of temperament traits from age 3 to 12. Thus, despite this being a period of greater plasticity and change relative to most other developmental periods (Posner & Rothbart, 2007; Rothbart & Bates, 2006), temperament shows substantial stability over the 9–10 year period spanning early childhood to early adolescence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Achenbach TM, McConaughy SH, & Howell CT (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. Psychological Bulletin, 101(2), 213. [PubMed: 3562706]

Bakermans-Kranenburg MJ, Van Ijzendoorn MH, & Juffer F (2003). Less is more: meta-analyses of sensitivity and attachment interventions in early childhood. Psychological bulletin, 129(2), 195–215. [PubMed: 12696839]

Capaldi DM, & Rothbart MK (1992). Development and validation of an early adolescent temperament measure. The Journal of Early Adolescence, 12(2), 153–173.

Carranza Carnicero JA, Pérez-López J, Del Carmen GS, & Martínez-Fuentes MT (2000). A longitudinal study of temperament in infancy: Stability and convergence of measures. European Journal of Personality, 14(1), 21–37.

Carranza JA, González-Salinas C, & Ato E (2013). A longitudinal study of temperament continuity through IBQ, TBAQ and CBQ. Infant Behavior and Development, 36(4), 749–761. [PubMed: 24036224]

Caspi A, Henry B, McGee RO, Moffitt TE, & Silva PA (1995). Temperamental origins of child and adolescent behavior problems: From age three to age fifteen. Child Development, 66(1), 55–68. [PubMed: 7497829]

Caspi A, & Roberts BW (2001). Personality development across the life course: The argument for change and continuity. Psychological Inquiry, 12(2), 49–66.

Caspi A, Roberts BW, & Shiner RL (2005). Personality development: Stability and change. Annual Review of Psychology, 56, 453–484.

Caspi A, & Silva PA (1995). Temperamental qualities at age three predict personality traits in young adulthood: Longitudinal evidence from a birth cohort. Child Development, 66, 486–498. [PubMed: 7750379]

Christensen A, Margolin G, & Sullaway M (1992). Interparental agreement on child behavior problems. Psychological Assessment, 4(4), 419.

Clark LA, McEwen JL, Collard LM, & Hickok LG (1993). Symptoms and traits of personality disorder: Two new methods for their assessment. Psychological Assessment, 5(1), 81–91.

Clark LA, & Watson D (2008). An organizing paradigm for trait psychology In John OP, & Robins RW (Eds.), Handbook of personality: Theory and research, 3rd *ed.* (pp. 265–286). Guilford Press New York.rd

De Pauw SSW, & Mervielde I (2010). Temperament, personality and developmental psychopathology: A review based on the conceptual dimensions underlying childhood traits. Child Psychiatry and Human Development, 41, 313–329. [PubMed: 20238477]

Dougherty LR, Bufferd SJ, Carlson GA, Dyson MW, Olino TM, & Klein DN (2011). Preschoolers' observed temperament and DSM-IV psychiatric disorders assessed with a parent diagnostic interview. Journal of Clinical Child and Adolescent Psychology, 40, 295–306. [PubMed: 21391025]

Durbin CE, Hayden EP, Klein DN, & Olino TM (2007). Stability of laboratory-assessed temperamental emotionality traits from ages 3 to 7. Emotion, 7(2), 388–399. [PubMed: 17516816]

Durbin CE, & Wilson S (2012). Convergent validity of and bias in maternal reports of child emotion. Psychological Assessment, 24(3), 647–660. [PubMed: 22149326]

Dyson MW, Olino TM, Durbin CE, Goldsmith HH, Bufferd SJ, Miller AR, & Klein DN (2015). The structural and rank-order stability of temperament in young children based on a laboratory-observational measure. Psychological Assessment, 27(4), 1388. [PubMed: 25894709]

Eisenberg N, Valiente C, Spinrad TL, Cumberland A, Liew J, Reiser M, ... & Losoya SH. (2009). Longitudinal relations of children's EC, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. Developmental Psychology, 45(4), 988–1008. [PubMed: 19586175]

Estabrook R, & Neale M (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. Multivariate Behavioral Research, 48(1), 1–27 [PubMed: 24049215]

Evans DE, & Rothbart MK (2007). Developing a model for adult temperament. Journal of Research in Personality, 41(4), 868–888.

Ferguson CJ (2010). A meta-analysis of normal and disordered personality across the life span. Journal of Personality and Social Psychology, 98(4), 659–667. [PubMed: 20307136]

Gagne JR, Van Hulle CA, Aksan N, Essex MJ, & Goldsmith HH (2011). Deriving childhood temperament measures from emotion-eliciting behavioral episodes: Scale construction and initial validation. Psychological Assessment, 23(2), 337. [PubMed: 21480723]

Goldsmith HH, Buss KA, & Lemery KS (1997). Toddler and childhood temperament: Expanded content, stronger genetic evidence, new evidence for the importance of environment. Developmental Psychology, 33(6), 891–905. [PubMed: 9383612]

Goldsmith HH.; Reilly J; Lemery KS; Longley S; Prescott A Technical report. Department of Psychology, University of Wisconsin-Madison; 1995 Laboratory Temperament Assessment Battery: Preschool Version.

Guerin DW, & Gottfried AW (1994). Developmental stability and change in parent reports of temperament: A ten-year longitudinal investigation from infancy through preadolescence. Merrill-Palmer Quarterly (1982-), 334–355

Hollingshead AB. Four factor index of social status Department of Sociology, Yale University; New Haven, CT: 1975 Unpublished manuscript.

Hopwood CJ, & Donnellan MB (2010). How should the internal structure of personality inventories be evaluated?. Personality and Social Psychology Review, 14, 332–346. [PubMed: 20435808]

Hoyle RH, & Panter AT (1995). Writing about structural equation models In Hoyle RH (Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 158–176). Thousand Oaks: Sage Publications

Jensen PS, Traylor J, Xenakis SN, & Davis H (1988). Child psychopathology rating scales and interrater agreement: I. Parents' gender and psychiatric symptoms. Journal of the American Academy of Child and Adolescent Psychiatry, 27, 442–450. [PubMed: 3182600]

John OP, Naumann LP, & Soto CJ (2008). The Big Five trait taxonomy: History, measurement, and conceptual perspectives In John OP, Robins RW, & Pervin LA (Eds.), Handbook of personality: Theory and Research (3rd ed., pp. 114–158). New York: Guilford Press.

Kenny DA, & Kashy DA (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112(1), 165–172.

Kiff CJ, Lengua LJ, & Zalewski M (2011). Nature and nurturing: Parenting in the context of child temperament. Clinical Child and Family Psychology Review, 14(3), 251–301. [PubMed: 21461681]

Klein DN, Dyson MW, Kujawa AJ, & Kotov R (2012). Temperament and internalizing disorders In Zentner M and Shiner R (Eds.), Handbook of Temperament (pp. 541–561). New York: Guilford Press.

Kline RB (1998). Software review: Software programs for structural equation modeling: Amos, EQS, and LISREL. Journal of Psychoeducational Assessment, 16(4), 343–364.

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

Kochanska G, Coy KC, Tjebkes TL, & Husarek SJ (1998). Individual differences in emotionality in infancy. Child Development, 69(2), 375–390. [PubMed: 9586213]

Komsi N, Räikkönen K, Heinonen K, Pesonen AK, Keskivaara P, Järvenpää AL, & Strandberg TE (2008). Continuity of father-rated temperament from infancy to middle childhood. Infant Behavior and Development, 31(2), 239–254. [PubMed: 18048099]

Komsi N, Räikkönen K, Pesonen AK, Heinonen K, Keskivaara P, Järvenpää AL, & Strandberg TE (2006). Continuity of temperament from infancy to middle childhood. Infant Behavior and Development, 29(4), 494–508. [PubMed: 17138302]

Kopala-Sibley DC, Dougherty LR, Dyson MW, Laptook RS, Olino TM, Bufferd SJ, & Klein DN (2017). Early childhood cortisol reactivity moderates the effects of parent–child relationship quality on the development of children's temperament in early childhood. Developmental Science, 20(3).

Kotelnikova Y, Olino TM, Klein DN, Kryski KR, & Hayden EP (2016). Higher-and lower-order factor analyses of the Children's Behavior Questionnaire in early and middle childhood. Psychological Assessment, 28(1), 92. [PubMed: 26029946]

Lemery KS, Goldsmith HH, Klinnert MD, & Mrazek DA (1999). Developmental models of infant and childhood temperament. Developmental Psychology, 35(1), 189. [PubMed: 9923474]

Linde JA, Stringer D, Simms LJ, & Clark LA (2013). The Schedule for Nonadaptive and Adaptive Personality for Youth (SNAP-Y) A New Measure for Assessing Adolescent Personality and Personality Pathology. Assessment, 20(4), 387–404. [PubMed: 23794180]

Lipscomb ST, Leve LD, Harold GT, Neiderhiser JM, Shaw DS, Ge X, & Reiss D (2011). Trajectories of parenting and child negative emotionality during infancy and toddlerhood: A longitudinal analysis. Child Development, 82(5), 1661–1675. [PubMed: 21883160]

Lo SL, Vroman LN, & Durbin CE (2015). Ecological validity of laboratory assessments of child temperament: Evidence from parent perspectives. Psychological Assessment, 27(1), 280–290. [PubMed: 25330108]

Majdandži M, & Van Den Boom DC (2007). Multimethod longitudinal assessment of temperament in early childhood. Journal of Personality, 75(1), 121–168. [PubMed: 17214594]

Mangelsdorf SC, Schoppe SJ, & Buur H (2000). The meaning of parental reports: A contextual approach to the study of temperament and behavior problems in childhood In Molfese V and Mofese DL (Eds.), Temperament and personality development across the life span (pp. 121–140). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Markon KE, Krueger RF, & Watson D (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. Journal of Personality and Social Psychology, 88, 139–157 [PubMed: 15631580]

Marsh HW, Hau KT, & Wen Z (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling, 11(3), 320–341.

Marsh HW, Scalas LF, & Nagengast B (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: traits, ephemeral artifacts, and stable response styles. Psychological Assessment, 22(2), 366. [PubMed: 20528064]

Martin RP (1994). Child temperament and common problems in schooling: Hypotheses about causal connections. Journal of School Psychology, 32(2), 119–134.

Melley AH, Oltmanns TF, & Turkheimer E (2002). The Schedule for Nonadaptive and Adaptive Personality (SNAP) Temporal stability and predictive validity of the diagnostic scales. Assessment, 9(2), 181–187. [PubMed: 12066833]

Moffitt TE, Arseneault L, Belsky D, Dickson N, Hancox RJ, Harrington H, ... & Sears MR (2011). A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences, 108(7), 2693–2698.

Muchinsky PM (1996). The correction for attenuation. Educational and Psychological Measurement, 56(1), 63–75.

Neppl TK, Donnellan MB, Scaramella LV, Widaman KF, Spilman SK, Ontai LL, & Conger RD (2010). Differential stability of temperament and personality from toddlerhood to middle childhood. Journal of Research in Personality, 44(3), 386–396. [PubMed: 20634996]

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

Olino TM, Klein DN, Dyson MW, Rose SA, & Durbin CE (2010). Temperamental emotionality in preschool-aged children and depressive disorders in parents: associations in a large community sample. Journal of Abnormal Psychology, 119(3), 468. [PubMed: 20677836]

Pedlow R, Sanson A, Prior M, & Oberklaid F (1993). Stability of maternally reported temperament from infancy to 8 years. Developmental Psychology, 29(6), 998.

Pesonen AK, Räikkönen K, Keskivaara P, & Keltikangas-Järvinen L (2003). Difficult temperament in childhood and adulthood: Continuity from maternal perceptions to self-ratings over 17 years. Personality and Individual Differences, 34(1), 19–31.

Plomin R, Caspi A, Pervin LA, & John OP (1999). Behavioral genetics and personality. Handbook of Personality: Theory and Research, 2, 251–276.

Podsakoff PM, MacKenzie SB, Lee JY, & Podsakoff NP (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of applied psychology, 88(5), 879–903. [PubMed: 14516251]

Posner MI, & Rothbart MK (2007). Research on attention networks as a model for the integration of psychological science. Annual Review of. Psychology, 58, 1–23.

Putnam SP, Gartstein MA, & Rothbart MK (2006). Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. Infant Behavior and Development, 29, 386–401. [PubMed: 17138293]

Putnam SP, Rothbart MK, & Gartstein MA (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. Infant and Child Development, 17(4), 387–405.

Ready RE, & Clark LA (2002). Correspondence of psychiatric patient and informant ratings of personality traits, temperament, and interpersonal problems. Psychological Assessment, 14(1), 39–49. [PubMed: 11911048]

Roberts BW, Caspi A, & Moffitt TE (2001). The kids are alright: growth and stability in personality development from adolescence to adulthood. Journal of Personality and Social Psychology, 81, 670–683. [PubMed: 11642353]

Roberts BW, & DelVecchio WF (2000). The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin, 126(1), 3. [PubMed: 10668348]

Roberts BW, & Mroczek D (2008). Personality trait change in adulthood. Current directions in psychological science, 17(1), 31–35. [PubMed: 19756219]

Rothbart MK (1981). Measurement of temperament in infancy. Child development, 569–578.

Rothbart MK, Ahadi SA, & Evans DE (2000a). Temperament and personality: origins and outcomes. Journal of Personality and Social Psychology, 78, 122. [PubMed: 10653510]

Rothbart MK, Ahadi SA, Hershey KL, & Fisher P (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. Child Development, 72(5), 1394–1408. [PubMed: 11699677]

Rothbart MK, & Bates JE (2006).Temperament. Handbook of Child Psychology. III:3.

Rothbart MK, Derryberry D, & Hershey K (2000b). Stability of temperament in childhood: Laboratory infant assessment to parent report at seven years In Molfese VJ, Molfese DL, & McCrae RR(Eds.), Temperament and Personality Development Across the Life Span (pp. 85–119).

Rothbart MK, Sheese BE, & Posner MI (2007). Executive attention and effortful control: Linking temperament, brain networks, and genes. Child Development Perspectives, 1(1), 2–7.

Sanson A, Hemphill SA, & Smart D (2004). Connections between temperament and social development: A review. Social Development, 13, 142–170.

Saudino KJ (2005). Behavioral genetics and child temperament. Journal of Developmental and Behavioral Pediatrics, 26(3), 214–223. [PubMed: 15956873]

Schafer JL, & Graham JW (2002). Missing data: our view of the state of the art. Psychological Methods, 7(2), 147. [PubMed: 12090408]

Shiner R, & Caspi A (2003). Personality differences in childhood and adolescence: Measurement, development, and consequences. Journal of Child Psychology and Psychiatry, 44(1), 2–32. [PubMed: 12553411]

Soper DS (2017). A-priori Sample Size Calculator for Structural Equation Models [Software]. Available from http://www.danielsoper.com/statcalc

Tackett JL, Krueger RF, Iacono WG, & McGue M (2008). Personality in middle childhood: A hierarchical structure and longitudinal connections with personality in late adolescence. Journal of Research in Personality, 42, 1456–1462. [PubMed: 19122851]

Tellegen A (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report.

Watson D (2000). Mood and Temperament. Guilford Press.

Watson D, & Clark LA (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. Journal of personality, 60(2), 441–476. [PubMed: 1635050]

Watson D, Clark LA, & Harkness AR (1994). Structures of personality and their relevance to psychopathology. Journal of Abnormal Psychology, 103, 18–31. [PubMed: 8040477]

Watson D, Gamez W, & Simms LJ (2005). Basic dimensions of temperament and their relation to anxiety and depression: A symptom-based perspective. Journal of Research in Personality, 39(1), 46–66.

Webster-Stratton C (1988). Mothers' and fathers' perceptions of child deviance: Roles of parent and child behaviors and parent adjustment. Journal of Consulting and Clinical Psychology, 56(6), 909–915. [PubMed: 3204202]

Westland JC (2010). Lower bounds on sample size in structural equation modeling. Electronic Commerce Research and Applications, 9(6), 476–487.

Youngstrom E, Izard C, & Ackerman B (1999). Dysphoria-related bias in maternal ratings of children. Journal of Consulting and Clinical Psychology, 67(6), 905–916. [PubMed: 10596512]

Zentner M, & Bates JE (2008). Child temperament: An integrative review of concepts, research programs, and measures. International Journal of Developmental Science, 2(1–2), 7–3.
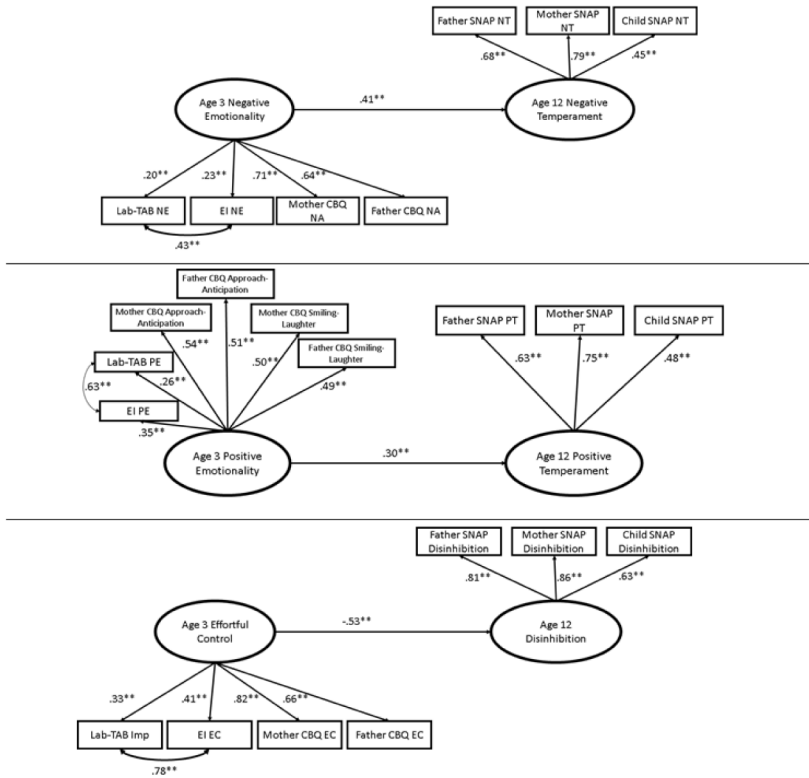
Figure 1.

** *p* < .01. Factor loadings of observed indicators and stability of temperament from age 3 to 12 years. All parameters are standardized estimates. Covariances on observed variable are covariances on the error terms of those variables. Error term on endogenous variables not depicted. Lab-TAB = Laboratory Temperament Assessment Battery, EC = Effortful Control, Imp = Impulsivity, EI = Experimenter Impressions, CBQ = Child Behavior Questionnaire, NA = Negative Affectivity, EC = Effortful Control, SNAP = Schedule for Non-adaptive and Adaptive Personality for Youth, NT = Negative Temperament, DIS = Disinhibition, PT = Positive Temperament, Mother = Mother-rated, Father = Father-rated, Child = Child-rated.
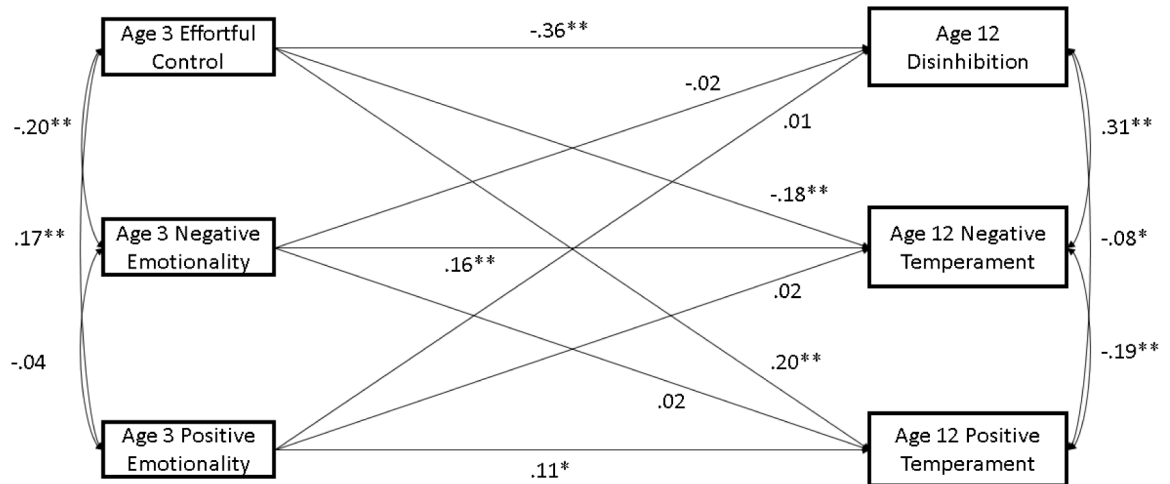
Figure 2.
Relationships between imputed latent Age 3 and 12 temperament traits. ** p < .01, *p < .05.
Note: Variables are rectangles as they were used as observed variables in the model, but
were derived by extracting latent scores from the latent models in Figure 1. All parameters
are standardized estimates. Parameters on curved arrows are correlations. Curved double-
headed arrows on endogenous variables are correlations on the error terms of those
variables.

Table 1.

Summary of variables used to measure theoretical constructs.

Variable and age	Theoretical Construct Measured	Sample Item
Lab-TAB negative emotionality (Age 3)	Negative emotionality	
Experimenter Impressions negative emotionality (Age 3)	Negative emotionality	
Lab-TAB positive emotionality	Positive emotionality	
Experimenter Impressions positive emotionality	Positive emotionality	
Lab-TAB Impulsivity (Age 3, reversed)	Effortful control	
Experimenter Impressions effortful control (Age 3)	Effortful control	
CBQ negative affectivity (Age 3)	Negative emotionality	"Cries sadly when a favorite toy gets lost or breakoken"
CBQ approach-anticipation (Age 3)	Positive emotionality	"Gets so worked up before an exciting event (s)he has trouble sitting still"
CBQ smiling-laughter (Age 3)	Positive emotionality	"Laughs a lot at jokes and silly happenings"
CBQ effortful control (age 3)	Effortful control	"Usually rushes into an activity without thinking about it" (reversed)
SNAP-Y positive temperament (Age 12)	Positive emotionality	"I get excited when I think about the future"
SNAP-Y negative temperament (Age 12)	Negative emotionality	" I often feel nervous and tense"
SNAP-Y disinhibition (Age 12)	Effortful control	"I never buy things on a whim or an impulse"

Note: Lab-TAB = Laboratory Temperament Assessment Battery, SNAP-Y = Schedule for Non-adaptive and Adaptive Personality for Youth, CBQ = Child Behavior Questionnaire.

Table 2.

Raw correlations between age 3 and 12 temperament variables.

Age 3 →	LT-NE	EI-NE	LT-PE	EI-PE	LT-Imp	EI-EC	Mot CBQ NA	Dad CBQ NA	Mom CBQ approach-anticipation	Dad CBQ approach-anticipation	Mom CBQ smiling-laughter	Dad CBQ smiling-laughter	Mom CBQ EC	Dad CBQ EC
Age 12														
N	559	537	559	537	559	537	518	401	518	401	518	401	518	401
Mother SNAP	0.07 (.15)	**0.16** (.003)	−0.07 (.18)	−0.02 (.72)	0.07 (.20)	−0.09 (.07)	**0.22** (<.001)	**0.18** (.001)	**0.11** (.04)	0.03 (.64)	−0.10 (.06)	−0.02 (.74)	**−0.23** (<.001)	**−0.16** (.04)
NT N = 426														
Father SNAP	0.01 (.90)	0.05 (.43)	−0.03 (.53)	−0.01 (.85)	0.07 (.43)	−0.04 (.46)	**0.17** (.001)	**0.34** (<.001)	0.07 (.23)	**0.13** (.03)	−0.06 (.24)	−0.05 (.43)	**−0.18** (.001)	**−0.25** (<.001)
NT N = 359														
Child SNAP	**0.10** (.04)	0.10 (.10)	0.01 (.83)	−0.01 (.95)	−0.01 (.28)	0.02 (.73)	0.08 (.12)	**0.11** (.05)	−0.06 (.27)	0.01 (.94)	0.01 (.90)	−0.06 (.29)	−0.07 (.18)	−0.06 (.26)
NT N = 423														
Mother SNAP	**−0.11** (.02)	−0.07 (.14)	**0.11** (.02)	**0.1** (.04)	0.01 (.95)	0.04 (.44)	−0.04 (.51)	−0.09 (.10)	**0.18** (<.001)	0.06 (.31)	**0.12** (.02)	0.08 (.13)	**0.24** (<.001)	**0.17** (.001)
PT N = 426														
Father SNAP	−0.03 (.52)	0.02 (.60)	**0.12** (.02)	0.05 (.37)	0.07 (.18)	−0.02 (.69)	0.01 (.88)	−0.07 (.20)	0.10 (.08)	**0.2** (<.001)	0.01 (.89)	**0.19** (.001)	0.07 (.19)	**0.24** (<.001)
PT N = 359														
Child SNAP	−0.02 (.65)	−0.08 (.22)	0.09 (.08)	0.08 (.13)	0.05 (.45)	0.02 (.63)	−0.01 (.85)	−0.09 (.13)	0.09 (.08)	0.05 (.33)	0.04 (.49)	**0.13** (.02)	**0.11** (.02)	**0.15** (.006)
PT N = 423														
Mother SNAP	0.07 (.19)	0.01 (.74)	−0.06 (.20)	−0.06 (.13)	**0.18** (<.001)	**−0.23** (<.001)	**0.12** (.02)	0.07 (.21)	0.06 (.19)	0.03 (.52)	**−0.14** (.004)	0.01 (.92)	**−0.40** (<.001)	**−0.25** (<.001)
DIS N = 426														
Father SNAP	0.05 (.38)	0.08 (.41)	−0.08 (.12)	−0.1 (.06)	**0.19** (.002)	**−0.2** (<.001)	−0.04 (.61)	0.1 (.07)	−0.01 (.98)	0.01 (.90)	−0.10 (.06)	−0.02 (.71)	**−0.33** (<.001)	**−0.39** (<.001)
DIS N = 359														
Child SNAP	0.08 (.10)	0.01 (.82)	−0.09 (.07)	0.08 (.10)	**0.16** (<.001)	**−0.18** (<.001)	0.02 (.55)	0.03 (.59)	−0.03 (.62)	0.03 (.47)	−0.03 (.56)	−0.02 (.65)	**−0.2** (<.001)	**−0.18** (.001)
DIS N = 423														

Note: Sample size (N) is included as this varied for each correlation. Values in parentheses under each correlation are p-values. Significant correlations (p < .05) are bolded. LT = Lab-TAB, EI = Experimenter Impressions during the Lab-TAB, NE =Negative Emotionality, PE = Positive Emotionality, EC = Effortful Control, CBQ = Child Behavior Questionnaire, NA = Negative Emotionality, SNAP = Schedule for Non-adaptive and Adaptive Personality for Youth, NT = Negative Temperament, PT = Positive Temperament, DIS = Disinhibition, Mom = Mother-rated, Dad = Father-rated, Child = Child-rated.

Table 3.

Unreliability-adjusted correlations between age 3 and 12 temperament variables.

Age 3	LT-NE	EI-NE	LT-PE	EI-PE	LT-Imp	EI-EC	Mot CBQ NA	Dad CBQ NA	Mom CBQ approach-anticipation	Dad CBQ approach-anticipation	Mom CBQ smiling-laughter	Dad CBQ smiling-laughter	Mom CBQ EC	Dad CBQ EC
Age 12														
N	559	537	559	537	559	537	518	401	518	401	518	401	518	401
Mother SNAP NT N = 426	.08	.19**	−.08	−.02	−.09	−.11+	.25**	.21**	.14*	.04	−.12+	−.02	−.26**	−.18**
Father SNAP NT N = 359	.01	.06	−.04	−.01	−.09	−.05	.19**	.39**	.09	.16*	−.07	−.06	−.20**	−.29**
Child SNAP NT N = 423	.12*	.12+	.01	−.01	.01	.02	.09	.12*	−.08	.00	.01	−.07	−.08	−.07
Mother SNAP PT N = 426	−.13*	−.09	.13*	.11*	.01	.05	−.05	−.10	.23**	.08	.15**	.10	.28*	.20**
Father SNAP PT N = 359	−.04	.02	.14*	.06	−.09	−.02	.01	−.08	.12+	.26**	.01	.24*	.08	.28**
Child SNAP PT N = 423	−.02	−.10	.10+	.09	−.07	.03	−.01	−.11	.11+	.07	.05	.17*	.13*	.18**
Mother SNAP DIS N = 426	.08	.01	−.07	−.07	−.23**	−.29**	.13*	.08	.08	.04	−.18**	.01	−.46**	−.30**
Father SNAP DIS N = 359	.06	.10	−.10	−.12+	−.25**	−.25**	−.05	.12+	−.01	.01	−.14+	−.03	−.38**	−.46**
Child SNAP DIS N = 423	.10	.01	−.11+	.09	−.21**	−.23**	.02	.04	−.04	.04	−.04	−.03	−.24**	−.22**

**
p < .01,

*
p <.05.

+
p < .10.

Note: Sample size (N) is included as this varied for each correlation. See Table 2 for exact p-values. LT = Lab-TAB, EI = Experimenter Impressions during the Lab-TAB, NE = Negative Emotionality, PE = Positive Emotionality, EC = Effortful Control, CBQ = Child Behavior Questionnaire, NA = Negative Emotionality, SNAP = Schedule for Non-adaptive and Adaptive Personality for Youth, NT = Negative Temperament, PT = Positive Temperament, DIS = Disinhibition, Mom = Mother-rated, Dad = Father-rated, Child = Child-rated.

Table 4.

Significant *r*-to-*z* zero-order correlation comparisons.

Correlation 1	Correlation 2	z	p
Negative Emotionality			
Father CBQ NA—Father SNAP-Y NT	Father CBQ NA—Mother SNAP-Y NT	2.45	.01
Father CBQ NA—Father SNAP-Y NT	Father CBQ NA –Child SNAP-Y NT	3.59	<.001
Effortful Control			
Mother CBQ EC -- Mother SNAP-Y Dis	Mother CBQ EC -- Child SNAP-Y Dis	−3.55	<.001
Mother CBQ EC -- Father SNAP-Y Dis	Mother CBQ EC -- Child SNAP-Y Dis	−2.09	.04
Father CBQ EC -- Father SNAP-Y Dis	Father CBQ EC -- Mother SNAP-Y Dis	−2.32	.02
Father CBQ EC -- Father SNAP-Y Dis	Father CBQ EC -- Child SNAP-Y Dis	−3.37	<.001
Mother CBQ EC -- Mother SNAP-Y Dis	LT-Imp -- Mother SNAP-Y Dis	−3.73	<.001
Mother CBQ EC -- Mother SNAP-Y Dis	LT-Imp -- Father SNAP-Y Dis	−3.28	.001
Father CBQ EC -- Father SNAP-Y Dis	LT-Imp -- Father SNAP-Y Dis	−3.03	.002
Mother CBQ EC -- Mother SNAP-Y Dis	EI EC -- Mother SNAP-Y Dis	−2.82	.005
Father CBQ EC -- Father SNAP-Y Dis	EI-EC -- Father SNAP-Y Dis	−3.03	.002

*** p < .01,

* p < .05.

Note: Each column indicates which correlation was used. Bolded correlation indicates which correlation was significantly larger. Positive emotionality correlation comparisons not listed here because none were significant. See Table 3 for correlations used to compute values in this Table. LT=Lab-TAB, EC = Effortful Control, Imp = Impulsivity, EI = Experimenter Impressions, CBQ = Child Behavior Questionnaire, NA = Negative Affectivity, SNAP-Y = Schedule for Non-adaptive and Adaptive Personality for Youth, NT = Negative Temperament, DIS = Disinhibition, Mother = Mother-rated, Father = Father-rated, Child = Child-rated.