Journal of Personality 81:2, April 2013

DOI: 10.1111/jopy.12000

Gender Differences in Young Children's Temperament Traits: Comparisons Across Observational and Parent-Report Methods

Thomas M. Olino,[1] C. Emily Durbin,[2] Daniel N. Klein,[3] Elizabeth P. Hayden,[4] and Margaret W. Dyson[3]

[1]University of Pittsburgh School of Medicine
[2]Michigan State University
[3]Stony Brook University
[4]University of Western Ontario

Abstract

Evidence supporting the continuity between child temperament and adult personality traits is accumulating. One important indicator of continuity is the presence of reliable gender differences in traits across the lifespan. A substantial literature demonstrates gender differences on certain adult personality traits and recent meta-analytic work on child samples suggests similar gender differences for some broad and narrow domains of temperament. However, most existing studies of children rely only on parent-report measures. The present study investigated gender differences in temperament traits assessed by laboratory observation, maternal-report, and paternal-report measures. Across three independent samples, behavioral observations, maternal-report, and paternal-report measures of temperament were collected on 463 boys and 402 girls. Across all three methods, girls demonstrated higher positive affect and fear and lower activity level than boys. For laboratory measures, girls demonstrated higher levels of sociability and lower levels of overall negative emotionality (NE), sadness, anger and impulsivity than boys. However, girls demonstrated higher levels of overall NE and sadness than boys when measured by maternal reports. Finally, girls demonstrated lower levels of sociability based on paternal reports. Results are discussed in relation to past meta-analytic work and developmental implications of the findings.

Contemporary models of temperament emphasize continuity of individual differences in emotion, motivation, and social behavior across the lifespan (Caspi & Shiner, 2006). Continuity may refer to similarity in the manifestation and structure of traits in children and adults or the extent to which traits have parallel correlates across the lifespan. There are well-replicated self-reported gender differences between men and women on higher- and lower-order dimensions of personality (Costa, Terracciano, & McCrae, 2001; Feingold, 1994; Lynn & Martin, 1997), raising questions of when these difference emerge during development and whether the magnitude of these differences changes over the course of development.

Meta-analyses of adult samples have explored gender differences on both higher- and lower-order personality traits. Feingold (1994) reported on gender differences on personality traits derived from multiple questionnaire measures, including those from the Five-Factor Model tradition (FFM; McCrae & Costa, 1987). Women demonstrated significantly higher levels of the higher-order trait of Extraversion, but no significant gender differences were found for Neuroticism, Agreeableness, Conscientiousness, or Openness to New Experience. For lower-order facets of the FFM, women were higher on anxiety (a facet of Neuroticism), gregariousness (a facet of Extraversion), and tendermindedness (a facet of Agreeableness). By contrast, men were higher on the assertiveness and self-esteem facets of Extraversion. In a more recent meta-analysis, Lynn and Martin (1997) found that for higher-order domains assessed by the Eysenck Personality Inventory (EPQ; Eysenck & Eysenck, 1975), women had higher levels of Neuroticism than men, while men scored higher than women on Extraversion and Psychoticism. Costa et al. (2001) examined studies using the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) and found that women scored higher than men on Neuroticism, Agreeableness, and components of Extraversion (warmth, gregariousness, positive emotions), and Openness

This work was supported by NIMH K01-MH092603 (TMO), RO1-MH069942 (DNK), and GCRC Grant M01-RR10710 to Stony Brook University from the National Center for Research Resources.

Direct correspondence to Thomas M. Olino, University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic, 3811 O'Hara St, Pittsburgh, PA, 15213. Email: olinotm@upmc.edu.

(openness to feelings). In contrast, men scored higher on the assertiveness and excitement seeking facets of Extraversion, and two facets of Openness (fantasy and openness to ideas).

In sum, these studies converge to suggest that the best replicated gender differences in personality traits in adulthood concern the two higher-order traits tapping emotionality. Compared to men, women describe themselves as higher on Neuroticism (particularly the facet of Anxiety). Findings for Extraversion varied by facet. Affiliative components of Extraversion, such as warmth and sociability, were higher in women than men (Costa et al., 2001). In contrast, measures of Assertiveness, as well as the EPQ Extraversion scale that includes much agentic content, were higher for men than women (Costa et al., 2001). Thus, in addition to differences in higher-order personality dimensions, lower-order aspects of certain traits may reveal additional effects, consistent with claims that lower-order constructs, rather than higher-order, provide greater power to detect individual differences (e.g., Costa et al., 2001). Finally, self-reports of the remaining traits, Agreeableness, Conscientiousness, and Openness, do not appear to differ across adult men and women.

Results of studies of adolescent samples largely parallel those of adults. In a birth cohort sample, Roberts, Caspi, and Moffitt (2001) found that 18-year-old females reported higher levels of constraint, the harm avoidance and stress reaction components of Neuroticism, and the affiliative aspects of Extraversion (positive emotionality, well-being) than males. By contrast, males were higher than females on aggression (an aspect of low Agreeableness), and agentic elements of Extraversion (achievement and social potency). Longitudinal developmental studies of adolescents complement these cross-sectional findings. In a meta-analysis exploring developmental change in self-reported personality traits, Roberts, Walton, and Viechtbauer (2006) found that gender did not moderate mean-level change through adolescence for any personality dimension explored, suggesting that most of the gender differences observed in adulthood are apparent in adolescence, implying that differences are early-emerging.

Thus, a better understanding of gender differences in personality/temperament traits in childhood is critical for understanding the developmental context in which gender differences are reliably observed in adults and adolescents first appear. Importantly, whereas the literatures on adults and adolescents largely employ self-report to assess these dimensions, the vast majority of studies of children use parent reports and a minority of studies use observational measures. Hence, our understanding of the emergence of gender differences in traits must also consider the influence of assessment method on the measurement of temperament.

In a meta-analysis on gender differences in child temperament, Else-Quest, Hyde, Goldsmith, and Van Hulle (2006) examined studies of children aged 3 months to 13 years. As our goal is to describe the evidence for the developmental continuity of personality, we summarize their results by emphasizing constructs related to the three higher-order dimensions that are identified in most models of adult personality (i.e., the "Big Three" of Extraversion/Positive Emotionality, Neuroticism/Negative Emotionality, and Constraint; Tellegen & Waller, 2008), as well as of child temperament (as exemplified by Rothbart's psychobiological model; Rothbart, Ahadi, Hershey, & Fisher, 2001), including the corresponding traits of Surgency/Positive Emotionality, Negative Affectivity, and Effortful Control. Thus, we focus on results from those studies that examined parent report measures of child temperament reflecting Rothbart's model and used the corresponding instruments, which captured a large minority of effect sizes in the meta-analysis.

Else-Quest and colleagues (2006) found no significant gender difference on Negative Affectivity ($d = -.06$). However, boys exhibited higher levels of broadband Surgency/PE than girls ($d = .55$) and girls demonstrated higher levels of Effortful Control ($d = -1.01$). For narrowband dimensions of Surgency, boys displayed higher levels of activity ($d = .23$), high-intensity pleasure ($d = .30$), and impulsivity ($d = .18$) than girls; no differences were found on approach ($d = -.04$), shyness ($d = -.03$), or smiling ($d = .01$). For narrowband dimensions of Negative Affectivity, boys displayed lower levels of fear ($d = -.12$) than girls. No significant differences were found for anger ($d = .04$), discomfort ($d = -.17$), distress to limits ($d = .01$), sadness ($d = -.10$), or soothability ($d = .05$). Among the Effortful Control narrowband dimensions, girls displayed higher levels of attentional focusing ($d = -.16$), attentional shifting ($d = -.16$), inhibitory control ($d = -.41$), low-intensity pleasure ($d = -.29$), and perceptual sensitivity ($d = -.38$) than boys.

The Else-Quest et al. (2006) meta-analysis suggests some continuity for some Big Three personality dimension gender differences in older adolescent, adult, and youth samples. For example, although boys and girls do not differ on overall NE as do adult men and women, girls are more fearful than boys. Other findings were less consistent with those from adult samples, but were similar to those found in adolescents. Specifically, girls had higher levels of several facets of Effortful Control, a trait corresponding most closely to Conscientiousness/Constraint. Although adult men and women do not differ on Conscientiousness, adolescent females are higher on Constraint (Roberts et al., 2001). Some gender differences in youngsters varied according to the developmental period in which they were assessed. However, the effects were small for each age group and no comparisons within these developmental periods were statistically significant.

Thus, it appears that some gender differences in temperament traits in adulthood are foreshadowed in childhood, but others do not reliably emerge until adolescence. However, differences between child and adolescent/adult samples could reflect methodological factors, rather than developmental effects. Most studies of gender differences in child temperament have relied on parent report measures of temperament. Far less is known about gender differences in child temperament when assessed by methods other than parent report. Else-

Quest et al. (2006) could not conduct moderator analyses to examine whether gender differences varied according to measurement approach as only eight studies relied on observational methods. Although useful for providing preliminary evidence, these studies are not definitive, as all but two had relatively modest sample sizes. Of the two with larger samples, both examined only a limited number of temperament dimensions, and one (Arcus & Kagan, 1995) did not provide enough information to compute effect sizes. The other large study reported that boys displayed higher activity level and greater difficult temperament, but lower distress to limitations and fear than girls across ages 3 to 5 (Zahn-Waxler, Schmitz, Fulker, Robinson, & Emde, 1996).

As few studies have used observational methods, it is difficult to determine whether gender differences are limited to a single methodology (i.e., parent reports), a particular reporter (i.e., mothers, whose reports are typically the principal source of parent questionnaires), or if they generalize to other assessment methods. Without such information, it is unclear whether gender differences on particular traits first appear in adolescence because of critical developmental transitions during that period, the effects of maturation or other factors that shape personality, or simply because self-report methods used in older samples are more sensitive to gender differences than parent-report measures typically used with child samples.

It is difficult to compare parent-report and observational studies of gender differences in child temperament traits due the well established fact that only modest associations of trait scores are found across methods (e.g., Gartstein & Marmion, 2008; Stifter, Willoughby, & Towe-Goodman, 2008). However, both assessment approaches have merits and are associated with important outcomes (Dougherty, Klein, Durbin, Hayden, & Olino, 2010; Hayden, Klein, & Durbin, 2005). Limited convergence across assessment methods suggests at least two possibilities for how gender differences may manifest. First, although parent report and observational methods exhibit only modest agreement on rank-ordering of trait levels of children, they may produce similar findings for mean-level differences between males and females. Alternatively, gender differences may vary (i.e., be moderated) by assessment method, a result that was not formally tested in Else-Quest et al. (2006). No published studies have evaluated these possibilities within the same data set. Thus, it is important to directly test whether patterns of gender differences are similar across assessment methods in order to make substantive interpretations about the presence and magnitude of gender differences in temperament traits in youngsters.

Here, we examined whether gender differences were evident in a broad range of temperament traits assessed using both laboratory-based observational measures and maternal- and paternal-reports. We used data collected from three community-based samples of preschool- and early elementary school-aged children. This developmental period is particularly important for identifying gender differences in temperamental traits, as temperament begins to stabilize around age 3 (e.g., Caspi & Shiner, 2006), suggesting that this may be the earliest age at which male-female differences can be reliably detected.

As previous findings suggest that nonsignificant differences on higher-order traits may mask significant differences on subordinate traits and findings for higher-order traits may not generalize to all lower-order dimensions, we included both broadband (i.e., higher-order) and narrowband (i.e., lower-order) temperament dimensions to discern the structural level at which gender differences on traits were most prominent. To directly compare gender differences assessed using multiple methods, we identified eight narrowband scales from Rothbart's Child Behavior Questionnaire (CBQ; Rothbart et al., 2001) that included similar content to behaviors coded from structured laboratory assessments of temperament. We selected three narrowband scales to represent a broad conceptualization of Positive Emotionality (PE; positive affect [PA], appetitive motivation, and sociability) and three to represent an overall Negative Emotionality (NE) construct (sadness, anger, and fear). We also examined impulsivity and activity level as two additional narrowband traits, but did not consider them as part of a higher-order trait, as their location within personality dimensions is controversial (Buss, Block, & Block, 1980; Eysenck, 1978). We selected these scales to maximize conceptual overlap with the Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith, Reilly, Lemery, Longley, & Prescott, 1995), which was developed using the same general theoretical framework as Rothbart's model.

We expected to replicate results from Else-Quest et al.'s (2006) meta-analysis for maternal reports, such that girls would have higher levels of fear and boys would have higher levels of activity and impulsivity. Although less information is available on paternal reports, we anticipated that results for fathers' reports of child temperament would be generally similar to those for mothers'. Finally, we had few strong predictions for gender differences on laboratory observational methods, for two reasons: (1) the available evidence regarding gender differences on child traits assessed via observation is inconsistent, and (2) as noted above, parent-report and observational measures typically demonstrate modest convergence (e.g., Seifer, Sameroff, Barrett, & Krafchuk, 1994), making it questionable to assume that results obtained via one method would be found for the other. However, given that sociocultural expectations regarding gender differences may have a stronger influence on parent-reports than on objective coding of child behavior in response to laboratory tasks (Seifer, 2003), we expected larger effect sizes for gender differences on parent-report than laboratory measures.

Methods

Data came from three studies of child temperament: the Stony Brook Temperament Study (SBTS), the Child Personality Development Project (CPDP), and the Northwestern Family Temperament Study (NFTS), yielding 865 child participants with one child included per family.

Table 1 Demographic Characteristics of the Three Samples

	SBTS	CPDP	NFTS	F/χ^2
Child Age (months)	42.24 (3.14)$_a$	43.20 (3.60)$_a$	56.38 (12.02)$_b$	365.46***
Child Sex, Male‡	302 (54.0)	53 (53.0)	108 (51.9)	.28
Child Race, Caucasian‡	487 (87.1)$_a$	89 (94.7)$_b$	158 (77.4)$_c$	18.37***
PPVT	102.82 (14.00)$_a$	103.47 (13.87)$_a$	106.62 (15.07)$_b$	6.73**
Maternal Age (years)	35.99 (4.44)$_a$	33.81 (4.10)$_b$	36.96 (4.89)$_c$	16.41***
Paternal Age (years)	38.27 (5.39)$_a$	36.73 (5.53)$_b$	38.72 (6.62)$_a$	4.19*
Maternal Employment‡	286 (51.2)$_a$	60 (56.6)$_{a,b}$	132 (64.0)$_b$	9.02*
Parent Marital Status‡	524 (93.7)	97 (97.0)	191 (92.7)	1.34
Maternal CBQ Completion‡	514 (91.9)$_a$	99 (99.0)$_b$	162 (77.6)$_c$	42.63***
Paternal CBQ Completion‡	399 (71.4)	82 (82.0)	156 (75.7)	5.24

Note. Table entries are *M (SD)*. Variables labeled as ‡ display *n (%)*. SBTS = Stony Brook Temperament Study; CPDP = Child Personality Development Project; NFTS = Northwestern Family Temperament Study; PPVT = Peabody Picture Vocabulary Test; CBQ = Child Behavior Questionnaire. Different subscripts reflect significant differences at $p < .05$.
*$p < .05$. **$p < .01$. ***$p < .001$.

The SBTS sample consisted of 559 three-year-old children and their parents from a suburban community in Long Island, New York. Participants were recruited through a commercial mailing list. Children who lived with at least one English-speaking biological parent and were free of significant medical conditions or developmental disabilities were included (Olino, Klein, Dyson, Rose, & Durbin, 2010). The CPDP sample consisted of 100 three-year-old children from Long Island, New York (Durbin, Klein, Hayden, Buckley, & Moerk, 2005). Children were recruited from a commercial mailing list (51.9%) and ads in local newspapers and preschools (48.1%). Participants obtained through the two methods did not differ on any of the child temperament variables used in this study. Participants in the NFTS sample ($N = 206$) were recruited from the greater Chicago area for a study of child temperament and were between the ages of 36 and 83 months. Participants were recruited through a commercial mailing list (38.1%), Internet, print, and radio ads (21.4%), referrals from community agencies (26.2%), and other approaches (e.g., word of mouth, 14.2%). Demographic characteristics for each study and comparisons between study samples are displayed in Table 1. Child gender distribution and percentage of children whose biological parents were currently married did not significantly differ between the samples. However, child age, race, Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) scores, maternal age, paternal age, and percentage of employed mothers differed between samples. Some differences reached statistical significance, but substantive implications were minimal. For example, differences in PPVT scores were at most a quarter of a standard deviation. The differences in child age and race across studies, particularly the NFTS relative to the CPDP and SBTS, bolster generalizability of the results.

Child Assessment Procedures

Parent Report Measures. For all studies, mothers and fathers completed the Child Behavior Questionnaire (CBQ; Rothbart et al., 2001). The CBQ is a widely used 195-item caregiver report measure of temperament for 3- to 7-year-old children. For the scales included in the present study, they have good internal consistency (mean $\alpha = .76$, range from .69 to .92 in the original publication), modest-to-strong inter-parental consistency (mean $r = .48$, range from .23 to .79), and moderate-to-strong test–retest reliability (mean $r = .67$, range from .55 to .79 for maternal reports and mean $r = .65$, from .58 to .76 for paternal reports) across a 2-year period (Rothbart et al., 2001). The CBQ derived scales are associated with concurrent home observations of temperament (Buckley, Klein, Durbin, Hayden, & Moerk, 2002) and prospectively associated with emotional and behavioral problem outcomes (Dougherty et al., 2010; Eisenberg, et al., 2003), thus showing both convergent and predictive validity.

In two of the samples (CPDP, SBTS), the parent who accompanied the child to the laboratory assessment (usually the mother) completed the CBQ during the lab visit, and questionnaires were sent home to the other parent to be completed and returned through postal mail. In the NFTS, both mothers and fathers completed the CBQ at home and returned them via postal mail. To maximize conceptual similarity of traits assessed by observational and parent-report measures, we focused on the Smiling/Laughter, Approach Anticipation, Shyness, Fear, Sadness, Anger, Impulsivity, and Activity Level scales from the CBQ. Internal consistency estimates are similar to those reported in Rothbart et al. (2001) and are presented for each scale for each individual study in Table 2. Higher-order PE was computed as the average of standardized values of Smiling/Laughter, Approach Anticipation, and Shyness (reverse scored). Higher-order NE was computed as the average of standardized values of Fear, Sadness, and Anger. As shown in Table 1, the percentage of mothers who completed the CBQ differed across the studies. Mothers from the CPDP had the highest completion percentage and mothers from the NFTS had the lowest. No differences were found in the percentage of fathers completing the CBQ between studies. Due to missing items on the CBQ, the actual Ns for each scale varied modestly.

Table 2 Temperament Dimension Characteristics

	SBTS				CPDP				NFTS			
	Laboratory		Mat.	Pat.	Laboratory		Mat.	Pat.	Laboratory		Mat.	Pat.
	ICC	α	α	α	ICC	α	α	α	ICC	α	α	α
PE												
PA	.92	.87	.73	.76	.94	.90	.82	.81	.90	.92	.71	.78
Engagement	.75	.68	.71	.71	.72	.56	.76	.65	.65	.69	.68	.62
Sociability	.83	.82	.92	.90	.93	.81	.92	.93	.93	.86	.92	.91
NE												
Fear	.64	.63	.73	.65	.66	.59	.74	.63	.66	.68	.76	.72
Sadness	.79	.81	.64	.66	.82	.67	.56	.60	.79	.74	.74	.61
Anger	.73	.68	.79	.76	.84	.75	.82	.83	.81	.74	.79	.81
Impulsivity	.74	.69	.76	.65	—	—	.70	.72	.70	.77	.77	.79
Activity	.84	.73	.76	.70	.75	.83	.71	.73	.94	.82	.79	.77

Note. SBTS = Stony Brook Temperament Study; CPDP = Child Personality Development Project (CPDP); NFTS = Northwestern Family Temperament Study; ICC = intra-class correlation; α = Cronbach's Alpha; PE = positive emotionality; PA = positive affectivity; NE = negative emotionality. ICCs are presented for directly observed behaviors (but not derived scores [i.e., PE and NE]). ICCs are based on 35 cases in the SBTS, 15 in the CPDP, and 27 in the NFTS. PA was indexed by the CBQ using the Smiling/Laughter scale. Sociability was indexed by the CBQ using the Shyness scale (reverse); Engagement was indexed by the CBQ using the Approach Anticipation scale.

Laboratory Assessment of Temperament. The laboratory batteries lasted approximately two hours, when children participated in standardized laboratory episodes with a female experimenter. Most episodes were from the Lab-TAB (Goldsmith et al., 1995); one (*Exploring New Objects*) was adapted from an original Lab-TAB episode, and two (*Making a T-shirt and Dress Up*) were developed by one of us (CED). Episodes were designed to elicit individual differences in temperament traits related to emotionality, behavioral engagement, and social behavior. The child took breaks between episodes to return to a baseline state before entering a new situation. Each task was videotaped through a one-way mirror and later coded. Although episodes are primed to elicit specific dimensions of temperament (indicated in parentheses below), all dimensions were rated in all episodes to provide indices of temperament across multiple contexts. The episodes are described below in the order that they were presented to the children in the SBTS and CPDP (numbers in brackets reflect the episode order for the NFTS).

Risk Room (fear; administered in SBTS and CPDP only). The episode allows children to explore a set of novel, ambiguous stimuli (e.g., a Halloween mask, a black box).

Tower of Patience (inhibitory control; interest; SBTS and CPDP only). The child and experimenter alternated turns in building a tower together. The experimenter took increasing delays before placing her block on the tower during each of her turns.

Making a T-shirt (PA; NFTS only [2]). The child decorated a T-shirt using fabric markers; he or she took the decorated T-shirt home as a gift.

Arc of Toys (PA; interest; anger; SBTS and CPDP only). The child played with toys for a five-minute period. The experimenter then asked the child to clean up the toys.

Disappointing Toy (sadness, anger; NFTS only [3]). The experimenter showed the child a picture of an unappealing toy and pictures of two appealing and asked the child which she or he would prefer. The experimenter left the room and returned with the nonpreferred toy. After 1 minute, an assistant entered with the preferred toy, and the child and experimenter played together for 3.5 min.

Stranger Approach (fear; SBTS, CPDP, and NFTS [4]). The child was left alone briefly in the assessment room while the experimenter left to look for toys. A male research accomplice entered the room and spoke to the child while walking closer.

Make That Car Go (PA, interest; SBTS and CPDP only). The child and experimenter raced remote controlled cars.

Dress Up (PA; NFTS only [5]). The child and experimenter played with costumes. The experimenter took a photograph of the child in his or her costume.

Transparent Box (persistence, interest, anger, sadness; SBTS, CPDP, and NFTS [6]). The experimenter locked an attractive toy in a transparent box. The child was then left alone with a set of keys to attempt to open the box. After a few minutes, the experimenter returned to the child and told them that she had left the wrong set of keys. The child was then encouraged to use the new keys to open the box and play with the toy.

Exploring New Objects (fear; SBTS, CPDP, and NFTS [1]). The child explored a set of novel and ambiguous stimuli (e.g., a mechanical spider, toy mice inside a pet carrier).

Pop-up Snakes (PA, interest; SBTS, CPDP, NFTS [9]). The child and experimenter surprised the child's mother with a can of potato chips that actually contained coiled snakes.

Perfect Circles (anger, sadness, persistence; SBTS, CPDP, and NFTS [8]). The experimenter repeatedly asked the child to draw a circle. Each attempt was mildly criticized. After about two minutes, the experimenter praised the child for his or her efforts.

Popping Bubbles (PA, interest; SBTS, CPDP, and NFTS [7]). The child and experimenter played with a bubble-shooting toy.

Snack Delay (inhibitory control; SBTS and CPDP only). The child was instructed to wait for the experimenter to ring a bell before eating a snack. The experimenter systematically increased the delay before ringing the bell.

Painting a Picture (interest; CPDP only). The child played with watercolor pencils and crayons.

Box Empty (anger, sadness; SBTS, CPDP, and NFTS [10]). The child was given an elaborately wrapped box, under the impression that a toy was inside. After the child discovered that the box was empty, the experimenter returned with several toys for the child to keep.

Laboratory Episode Coding Procedures

Affective Codes in SBTS and NFTS. Each display of facial, bodily and vocal positive affect, fear, sadness, and anger in each episode was rated on a three-point scale (low, moderate, high intensity). Weighted sums (low intensity = 1; moderate = 2; high = 3) were computed separately within each channel (facial, bodily, vocal) across the episodes, standardized, and then summed across the channels to derive total scores for positive affect, fear, sadness, and anger.

Affective Codes in CPDP. Discrete emotions (positive affect, anger, sadness, and fear) were assessed by coding facial, vocal, and bodily indicators during each episode. Each episode was scored (from 0–3 or 0–4) based on the number and intensity of affective displays. These indicators were averaged to produce composite variables for each emotion.

Behavioral Codes. Ratings of additional dimensions of child behavior were made via single global ratings (scored 0–3) based on all behaviors relevant to each dimension during that episode. Ratings of engagement were based on how invested and absorbed the child appeared in play. Sociability ratings were based on the quality and quantity of the child's attempts to engage and interact with the experimenter and, to a lesser extent, the parent. Activity level ratings were based on the quantity and quality of movement during each episode and the amount of vigor exhibited in manipulation of the stimuli. Impulsivity ratings (available in the SBTS and NFTS) ranged from low (deliberate, planful) to high (lacking inhibitory control). Internal consistency (α) and inter-rater reliability (intra-class correlations; ICC) for all narrowband temperament traits are displayed in Table 2.

Consistent with other studies using observational methods, a subset of cases were coded for reliability. For the SBTS, there were a total of 35 raters, including TMO, MWD, three graduate students, and 30 undergraduate students. TMO, MWD, and the three graduate students were trained on all episodes, and coded a random subset of children. Undergraduate research assistants were trained to code one or two episodes, which was done to reduce coder biases across episodes within the same participant. For the NFTS, one of the authors (CED), three graduate students, and approximately 26 undergraduate research assistants served as the raters. Similar processes were relied on in the CPDP with three of the authors (CED, EPH, and TMO) coding a portion of those observations along with another graduate student and approximately twelve undergraduate students. ICCs were estimated using two-way random effects for a single rater in each study (Shrout & Fleiss, 1979). Following guidelines from Shrout (1998), inter-rater reliability for all temperament dimensions was strong. Internal consistency and inter-rater reliability statistics were strikingly similar across all three studies. Observational codes were standardized for all analyses.

Data Analysis

First, we present independent sample *t*-tests comparing boys and girls on each temperament dimension separately for each method and rater. These provide readily interpretable effect size estimates, reported as Cohen's *d* (Cohen, 1992). Second, we compare the magnitude of gender differences across assessment method. Individual temperament scores from laboratory observations and maternal and paternal reports were standardized within each study, yielding three trait scores for each participant. Multilevel models (MLM) were estimated to examine whether gender differences in dimensions of temperament varied according to assessment methodology. All analyses were conducted in Mplus 6.11 (Muthén & Muthén, 1998–2010) using the TWOLEVEL and COMPLEX options. These options were specified to identify multiple levels of clustering: multiple assessments of the same dimension were nested within individual child participants and participants were nested within studies. As there were three assessment methods, we created a set of dummy-coded variables to predict temperament scores at the within-subject level. Dummy codes were constructed such that laboratory observation was the reference assessment methodology. Interaction effects were computed as the cross-product of gender and assessment. Interactions were interpreted only when the set (i.e., maternal report [vs. laboratory observation] × child sex; paternal report [vs. laboratory observation] × child sex) provided a significant improvement in model fit, as indexed by a log-likelihood difference test (−2LL). Models were computed using robust maximum likelihood estimation methods to accommodate missing data. Analyses included child age (in months) as a covariate.

Results

Convergent and Discriminant Associations

To assess convergence within traits across methods, we computed the average correlation for lower-order traits within the same broad temperament domain across laboratory and parent-report assessment methods. Consistent with the existing literature, we found modest associations between laboratory/observational and parent-report measures of traits. For lower-

order PE traits, the average correlation was .17; for lower-order NE traits, the average correlation was .10, and for lower-order constraint traits (i.e., impulsivity and activity level), the average correlation was .22. We also computed the average correlation across traits and across methods to index discriminant validity of traits. Across assessment methods, the average correlation between PE and NE was −.01; between PE and Constraint was .13; and between NE and Constraint was .03. Thus, associations within traits across methods were somewhat larger than associations across traits and across methods. Similar, albeit even more modest, patterns of convergent and divergent associations were found at the facet level.

Comparisons Between Boys and Girls for Each Assessment Method

Initial comparisons between boys and girls were conducted using independent samples *t*-tests (naive to complex sampling methods); means and standard deviations were used to compute Cohen's *d* (Table 3) for each temperament dimension. Boys and girls did not significantly differ on broadband PE, or its lower-order traits of PA or engagement using laboratory, maternal report, or paternal report methods; *d*s for each were small. Girls demonstrated significantly higher levels of sociability than boys as assessed using laboratory methods (a small effect size), but for maternal and paternal report methods, gender differences were small and nonsignificant.

For overall NE, no significant gender differences were found using laboratory observation or paternal report methods. However, mothers reported significantly higher NE in girls than boys (a small effect). A similar pattern was observed for sadness; no significant gender differences were found based on laboratory observation or paternal report methods, but there was a small, significant effect of child gender on mothers' reports of sadness. Across all three methods, girls demonstrated significantly higher levels of fear than boys, albeit a small effect. For anger, no significant gender differences were found using maternal or paternal reports. However, boys demonstrated significantly higher levels of anger than girls when assessed using laboratory measures; the effect size was small.

For impulsivity, no significant gender differences were found using maternal or paternal reports. However, boys were significantly higher than girls on laboratory-rated impulsivity, with a medium size effect ($d = .72$). Finally, boys were significantly higher in activity level than girls across all three methods; for each, the effect was small. Inconsistent with our prediction, the magnitude of gender differences across traits was larger for laboratory tasks (mean absolute value of $d = .24$) than for maternal or paternal reports (mean absolute value of $d = .14$ for both), suggesting that structured lab tasks were somewhat more sensitive in detecting gender differences than were parental reports. In addition, for several traits, gender differences were in opposite directions for different methods (e.g., $d = -.19$ for laboratory-assessed sociability versus .16 for paternal reports).

Comparison of Gender Differences as Assessed by Multiple Methods

In order to compare the effect sizes for gender across methods for each trait, we conducted MLM analyses. Scores were standardized within each method and sample. Models included main effects of child gender and assessment method (as a set of dummy codes) and interaction terms between child gender and assessment method dummy codes.

For overall PE and engagement, we found no significant effects of gender or interactions (Table 4). There was a significant main effect of gender and paternal report (vs. laboratory observation) for PA. The gender effect revealed that girls demonstrated higher levels of PA than boys. There were significant

Table 3 Means, *SD*s, and *d* Estimates for Boys and Girls for the Aggregated Sample

	Laboratory Observation			Maternal Report			Paternal Report		
	Boys M (SD)	Girls M (SD)	d	Boys M (SD)	Girls M (SD)	d	Boys M (SD)	Girls M (SD)	d
PE	−.05 (1.01)	.06 (.99)	.17	5.21 (.60)	5.23 (.63)	.03	5.15 (.59)	5.11 (.56)	−.07
PA	−.04 (.98)	.05 (1.02)	.13	5.91 (.62)	5.95 (.67)	.08	5.77 (.71)	5.76 (.69)	−.01
Engagement	−.03 (1.04)	.03 (.95)	.08	5.23 (.62)	5.29 (.61)	.11	5.04 (.59)	5.07 (.57)	.05
Sociability	−.06 (.98)	.07 (1.02)	.19*	4.49 (1.27)	4.45 (1.23)	−.04	4.63 (1.08)	4.51 (1.13)	−.16
NE	.02 (.99)	−.02 (1.01)	−.06	4.06 (.62)	4.18 (.63)	.21**	3.97 (.59)	4.03 (.55)	.10
Fear	−.11 (.99)	.12 (1.00)	.32**	3.89 (.95)	4.04 (.89)	.22*	3.74 (.84)	3.90 (.79)	.25*
Sadness	.02 (.93)	−.02 (1.07)	−.06	3.82 (.70)	4.00 (.70)	.30***	3.71 (.67)	3.80 (.62)	.16
Anger	.11 (1.07)	−.13 (.89)	−.34**	4.48 (.83)	4.51 (.85)	.05	4.46 (.79)	4.37 (.77)	−.15
Impulsivity	.23 (1.05)	−.27 (.87)	−.72***	4.68 (.78)	4.60 (.74)	−.14	4.67 (.65)	4.59 (.65)	−.15
Activity	.10 (.99)	−.12 (.99)	−.32**	5.02 (.79)	4.86 (.79)	−.26**	5.02 (.66)	4.87 (.70)	−.25**

Note. Laboratory Observations: Boys $n = 462$ (except for Impulsivity; $n = 409$); Girls $n = 401$–403 (except for Impulsivity; $n = 356$); Maternal Report: Boys $n = 424$–428; Girls $n = 350$–357; Paternal Report: Boys $n = 349$–351; Girls $n = 288$. PE = positive emotionality; PA = positive affectivity; NE = negative emotionality.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4 Multilevel Model Regression Parameters to Examine Gender Differences, Assessment Methodology Differences, and Gender × Methodology Interactions for Each Temperament Dimension

	G v. B	M v. L	D v. L	−2LL	Interactions
PE	.11 (.07)	.03 (.04)	.08 (.02)***	4.40	
PA	.09 (.04)*	.01 (.04)	.05 (.02)*	3.19	
Engagement	.06 (.137)	−.01 (.08)	.02 (.03)	0.30	
Sociability	.13 (.05)**	.08 (.02)**	.12 (.03)***	29.60***	Sex*M v. L: −.17 (.05)**; Sex*D v. L: −.25 (.04)***
NE	−.06 (.01)***	−.12 (.02)***	−.09 (.02)***	15.39***	Sex*M v. L: .25 (.04)***; Sex*D v. L: .20 (.07)**
Fear	.20 (.03)***	.03 (.01)*	.02 (.04)	0.42	
Sadness	−.05 (.01)***	−.13 (.03)***	−.08 (.05)	9.54**	Sex*M v. L: .30 (.06)***; Sex*D v. L: .18 (.11)
Anger	−.25 (.07)**	−.14 (.04)*	−.07 (.03)	29.67***	Sex*M v. L: .28 (.07)***; Sex*D v. L: .13 (.06)*
Impulsivity	−.49 (.02)***	−.17 (.02)***	−.17 (.03)***	112.61***	Sex*M v. L: .36 (.04)***; Sex*D v. L: .34 (.08)***
Activity	−.22 (.08)**	−.01 (.05)	−.00 (.06)	0.00	

Note. G = girls, B = boys; M = maternal report, D = paternal report, L = laboratory observation; PE = positive emotionality; PA = positive affectivity; NE = negative emotionality. −2LL = log-likelihood difference test for significance of the block of interaction terms. Interactions are displayed when the block of interactions were significant. All associations are adjusted for child age and include study as an additional layer of nonindependence.
*$p < .05$. **$p < .01$. ***$p < .001$.

main effects of gender and assessment method on sociability; however, these were qualified by significant gender × assessment method interactions. Follow-up analyses revealed that girls demonstrated significantly higher levels of sociability than boys for the laboratory assessment, boys had higher levels of sociability than girls based on paternal reports, and no differences were found on maternal reports.

For overall NE, significant main effects were found for gender and assessment method. However, these were qualified by significant gender × assessment method interactions. Follow-up analyses showed that boys demonstrated higher levels of NE than girls using observational methods; girls demonstrated significantly higher levels of NE than boys using maternal reports; and no interaction was found between child gender and paternal reports. For fear, the main effect of gender was significant, with girls demonstrating significantly higher levels of fear than boys. However, the gender × assessment method interaction was nonsignificant. For sadness, the main effects of gender and maternal report were significant; however, these were qualified by a gender × assessment method interaction. Girls demonstrated significantly lower levels of sadness than boys during the laboratory assessment; higher levels of sadness based on maternal reports; and no difference based on paternal reports. For anger, the main effects of gender and maternal report were both significant; however, these were qualified by a gender × assessment method interaction. Boys demonstrated significantly higher levels of anger than girls based on laboratory observations, but not on maternal or paternal reports.

For impulsivity, the main effect of gender was significant, but was qualified by the significant gender × assessment method interaction. Follow-up analyses found that boys demonstrated significantly higher levels of impulsivity than girls using laboratory and maternal report methods. However, the magnitude of this difference was smaller for maternal reports than laboratory observations. The difference for paternal report was significant at the level of a trend ($p = .059$), with boys being more impulsive than girls. For activity level, the main effect of gender was significant, such that boys demonstrated higher activity than girls. The main effect of assessment method and the gender × assessment method interaction were nonsignificant.

Last, we conducted a set of models examining whether the results differed across the three samples. These models included the main effects of gender, assessment method, and study, and all two- and three-way interactions and age as a covariate. The three-way interactions did not reach significance for any temperament dimension, indicating similar effects across samples that varied in demographic characteristics, including child age and race/ethnicity.

Discussion

Previous meta-analytic work examining gender differences in personality in adolescents and adults (Costa et al., 2001; Feingold, 1994; Lynn & Martin, 1997) has relied on self-reports and in child temperament (Else-Quest et al., 2006) has relied almost exclusively on parent-report questionnaires. The present study adds to this literature in two important ways. First, we report on the presence and magnitude of gender differences in child temperament traits as assessed by structured laboratory tasks, a method that is increasingly being employed by developmental scientists to understand individual differences in temperament, emotion, and social behavior (e.g., Hane, Fox, Henderson, & Marshall, 2008; Kochanska, Aksan, & Carlson, 2005). Second, we directly compared gender difference estimates from laboratory tasks to those from parent reports by both mothers and fathers within the same children. Thus, we were able to test whether the results of previous results generalize across assessment approaches, and to directly examine the relative magnitude of gender differences across methods. We focused on traits identified in major

models of both adult personality and child temperament to explore whether gender differences identified in the adult literature were also evident in young children. Importantly, we integrated data from three independent samples, and found no evidence that the results varied across samples.

The results of this study suggest that gender differences in the major temperament traits are smaller in young children than in samples of adults or adolescents. Moreover, the presence, direction, and magnitude of gender differences varied according to temperament dimension and method of assessment. Consistent with the Else-Quest et al. (2006) meta-analysis, no significant gender difference was found for engagement across laboratory assessment, maternal reports, or paternal reports. This cross-method consistency suggests that boys and girls have similar levels of engagement early in development, in contrast to the significant gender differences on related traits (e.g., ambition, endurance, and achievement; Watson & Clark, 1997) evident in adults. For positive mood, we found a small, but significant gender difference consistent with previous adult literature finding that women demonstrate higher levels of related traits (e.g., smiling; LaFrance, Hecht, & Paluck, 2003). For the third component of PE we explored—sociability—Else-Quest et al. reported that boys and girls did not significantly differ. Our results supported this finding for maternal reports. However, similar to results in adults (Costa et al., 2001), we found that girls demonstrated higher levels of sociability than boys when assessed using laboratory observations. It is possible that girls' greater verbal facility may have provided more salient observable examples of sociability to coders. Alternatively, the laboratory context, where children interacted with an unfamiliar female experimenter, may have provided a contextual press that elicited greater sociability in girls than in boys. In contrast, based on paternal reports, boys demonstrated more sociability than girls. Perhaps fathers are more actively engaged with their sons more than daughters, which may be revealed in their ratings of sociability. Interestingly, this was the only gender difference reported by fathers that was not also reported by mothers.

Although there is a moderate gender difference in adult samples for NE, such that women have higher levels than men, Else-Quest et al. (2006) did not find a significant gender difference for the higher-order NE dimension in young children. In contrast, we found differences for overall NE and narrow-band facets of NE for observational measures and maternal reports. In the lab, we found that boys demonstrated modestly greater overall NE than girls. However, as assessed by maternal reports, girls had higher levels of NE. Similar to prior studies of children (Else-Quest et al., 2006) and adults (Feingold, 1994), we found that girls demonstrated significantly higher levels of fear than boys across laboratory assessment, maternal report, and paternal report. This suggests that gender differences in fearfulness are evident quite early in development. Consistent with Else-Quest et al., we found that anger did not differ between boys and girls when assessed using maternal or paternal reports. However, boys demonstrated higher levels of anger in response to the laboratory tasks than did girls. Inconsistent with Else-Quest et al.'s findings, we found that boys demonstrated higher levels of sadness than girls during laboratory observations; girls had higher levels of sadness than boys according to maternal reports; and no differences were evident for this trait when using paternal reports. Thus, gender differences on anger and sadness were less consistent across methods than the differences for fear.

Regarding the final two dimensions of temperament we considered, we found that boys demonstrated higher levels of activity than girls across laboratory assessment, maternal report, and paternal report. This converges with the results reported by Else-Quest et al. (2006) for maternal reports. We also found significant gender differences on impulsivity for laboratory observations and maternal reports (and paternal reports at a trend level), with gender differences being more pronounced for laboratory measures than parent reports. This suggests that laboratory observations are particularly sensitive to gender differences in impulsivity, while differences in activity level are evident across all methods.

Differences across methods in their estimation of the magnitude of gender differences in temperament traits could emerge for multiple reasons. First, some methods may be more strongly influenced by gender norms for particular behaviors. For example, parents' ratings may be more strongly influenced by their expectations for general child behavior, rather than their own child's behavior. To the extent that these expectations validly represent gender differences on the trait, these may manipulate the size of the observed gender differences. Second, certain methods may be particularly sensitive to gender differences to the extent that they draw upon rich information regarding the construct. For example, parents may have better knowledge of traits expressed at a low base rate in the home than would raters who observe the child in a single visit to the home. By contrast, laboratory tasks are designed to elicit manifestations of traits in response to structured stimuli, which may be more sensitive to the rank-ordering of children. However, as effect sizes tended to be most modest for paternal reports it appear that fathers were least sensitive to gender differences.

Our comparisons of the broadband dimensions are not strictly comparable to those of Else-Quest and colleagues (2006). Our broadband PE construct included PA, interest, and sociability, while Rothbart's model (Rothbart et al., 2001) includes activity level and impulsivity on the Surgency/PE factor. This difference in operationalization may account for the gender difference favoring boys on Surgency reported by Else-Quest et al. For the approach, smiling/laughter, anticipation, and shyness scales described in Else-Quest et al., the average *d* was −.02. In the present study, the average *d* for PE across all methods was −.04. Thus, when considering strictly analogous constructs, the results are quite comparable. Similarly, our broadband NE construct included fear, sadness, and anger, while Rothbart's model includes additional indices of NE (e.g., discomfort, low soothability). For the sadness, anger, and fear scales in Else-Quest et al.'s study, the average *d* was

−.06. In the present study, across all methods, this was −.08. Thus, the results are again quite comparable. Therefore, the present findings suggest that gender differences in broadband dimensions of temperament assessed using different methods are quite comparable, although gender differences across methods are evident at the lower facet level.

In addition to addressing questions of gender differences, the present findings have important implications for future work exploring the processes of socialization and development of temperament and examinations of relationships between temperament traits and later developmental outcomes. Common to this work is research comparing predictive associations using multiple methods. For example, the relationship between gender and social functioning may be (partially) mediated by dimensions of temperament; for example, gender differences in attentional control may partially explain differences between boys and girls in their later social skills and sociometric ratings (e.g., Eisenberg et al., 1993). The results of these questions will likely differ depending upon how the trait in question is measured. Some traits differ across genders regardless of how they are measured (e.g., fear); thus, one would expect to find similar evidence regardless of assessment strategy for these traits. However, other traits (such as sadness or anger proneness) do not demonstrate equivalent gender differences across methods of assessment. Thus, it may be particularly important to include multiple measures of these traits when testing their associations with subsequent outcomes.

This study had several merits, including the use of three independent community-based samples, which produced a large total sample size and enhanced the generalizability of the findings; use of laboratory observational methods; and assessment of multiple traits at both broad and narrow levels. However, some limitations should also be acknowledged. First, observations were made in the context of laboratory tasks, which may limit the generalizability of the findings to other times and contexts. However, previous work from our group has found robust correlations between structured laboratory and naturalistic home observations and moderate stability over time (Buckley et al., 2002; Durbin, Hayden, Klein, & Olino, 2007). Second, our selection of parent-report measures was constrained to traits similar to those measured in the laboratory. Thus, we included a relatively small number of parent-report scales in our analyses. Third, there were only modest convergent associations across methods, and the difference between the convergent and discriminant associations were also modest. This was not unexpected, as similar findings have been reported numerous times in the child temperament literature. To the extent that traits assessed by different approaches tap distinct constructs, it complicates the interpretation of gender × assessment method interactions. Finally, while we argue that these findings have relevance for development, we used a cross-sectional design. Thus, it will be important to examine similar questions beginning earlier in development and following the same participants longitudinally to identify the mechanisms responsible for gender differences.

The effect sizes for gender differences reported in this paper and previous meta-analytic studies of parent-reported child temperament traits are uniformly smaller than those in the adult personality literature. Thus, although we found some evidence that adult gender differences are replicated for some traits in early childhood, it appears that these differences increase in magnitude between childhood and adolescence/early adulthood. This may be due to different mechanisms underlying personality change in boys or girls or to differential impact of the same mechanisms or contexts across genders. Further research mapping the development of temperament traits through early and middle childhood to early adolescence will be critical for understanding when gender differences emergence and the mechanisms underlying such changes. Finally, our results indicate that gender differences in temperament are moderated by the method by which traits are assessed. There are at least two important implications of this finding. First, for studies of children, researchers should use caution when interpreting gender differences assessed using a single method, and future studies should investigate the reasons for cross-method differences. Second, investigations of measurement influences on gender differences in adult personality should be explored. Most of the literature on this topic has relied on self-report methods (with some exceptions, e.g., Spinath, Angleitner, Borkenau, Riemann, & Wolf, 2002). Explorations of gender differences on traits using other assessment approaches (e.g., peer/significant other report, laboratory assessments) may suggest a different pattern of findings than those in the extensive self-report literature in adults, or further substantiate findings from self-report by demonstrating that they are robust across method.

References

Arcus, D., & Kagan, J. (1995). Temperament and craniofacial variation in the first two years. *Child Development*, **66**, 1529–1540.

Buckley, M. E., Klein, D. N., Durbin, C. E., Hayden, E. P., & Moerk, K. C. (2002). Development and validation of a *q*-sort procedure to assess temperament and behavior in preschool-age children. *Journal of Clinical Child and Adolescent Psychology*, **31**, 525–539.

Buss, D. M., Block, J. H., & Block, J. (1980). Preschool activity level: Personality correlates and developmental implications. *Child Development*, **51**, 401–408.

Caspi, A., & Shiner, R. L. (2006). Personality development. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 3, Social, emotional, and personality development* (pp. 265–286). Hoboken, NJ: John Wiley.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155–159.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising

findings. *Journal of Personality and Social Psychology*, **81**, 322–331.

Dougherty, L. R., Klein, D. N., Durbin, C. E., Hayden, E. P., & Olino, T. M. (2010). Temperamental positive and negative emotionality and children's depressive symptoms: A longitudinal prospective study from age three to age ten. *Journal of Social and Clinical Psychology*, **29**, 462–488.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test (3rd ed.)*. Circle Pines, MN: American Guidance Service.

Durbin, C. E., Hayden, E. P., Klein, D. N., & Olino, T. M. (2007). Stability of laboratory-assessed temperamental emotionality traits from ages 3 to 7. *Emotion*, **7**, 388–399.

Durbin, C. E., Klein, D. N., Hayden, E. P., Buckley, M. E., & Moerk, K. C. (2005). Temperamental emotionality in preschoolers and parental mood disorders. *Journal of Abnormal Psychology*, **114**, 28–37.

Eisenberg, N., Fabes, R. A., Bernzweig, J., Karbon, M., Poulin, R., & Hanish, L. (1993). The relations of emotionality and regulation to preschoolers' social skills and sociometric status. *Child Development*, **64**, 1418–1438.

Eisenberg, N., Valiente, C., Fabes, R. A., Smith, C. L., Reiser, M., Shepard, S. A., . . . Cumberland, A.J. (2003). The relations of effortful control and ego control to children's resiliency and social functioning. *Developmental Psychology*, **39**, 761–776.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, **132**, 33–72.

Eysenck, H. J. (1978). Superfactors P, E and N in a comprehensive factor space. *Multivariate Behavioral Research*, **13**, 475–481.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (junior and adult)*. London: Hodder & Stoughton.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, **116**, 429–429.

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

Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., & Prescott, A. (1995). *Laboratory Temperament Assessment Battery: Preschool version*. Unpublished manuscript.

Hane, A. A., Fox, N. A., Henderson, H. A., & Marshall, P. J. (2008). Behavioral reactivity and approach-withdrawal bias in infancy. *Developmental Psychology*, **44**, 1491–1496.

Hayden, E. P., Klein, D. N., & Durbin, C. E. (2005). Parent reports and laboratory assessments of child temperament: A comparison of their associations with risk for depression and externalizing disorders. *Journal of Psychopathology and Behavioral Assessment*, **27**, 89–100.

Kochanska, G., Aksan, N., & Carlson, J. J. (2005). Temperament, relationships, and young children's receptive cooperation with their parents. *Developmental Psychology*, **41**, 648–660.

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: A meta-analysis of sex differences in smiling. *Psychological Bulletin*, **129**, 305–334.

Lynn, R., & Martin, T. (1997). Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *Journal of Social Psychology*, **137**, 369–373.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, **52**, 81–90.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide (6th Ed.)*. Los Angeles, CA: Muthén & Muthén.

Olino, T. M., Klein, D. N., Dyson, M. W., Rose, S. A., & Durbin, C. E. (2010). Temperamental emotionality in preschool-aged children and depressive disorders in parents: Associations in a large community sample. *Journal of Abnormal Psychology*, **119**, 468–478.

Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of Personality and Social Psychology*, **81**, 670–683.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, **132**, 1–25.

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

Seifer, R. (2003). Twin studies, biases of parents, and biases of researchers. *Infant Behavior and Development*, **26**, 115–117.

Seifer, R., Sameroff, A. J., Barrett, L. C., & Krafchuk, E. (1994). Infant temperament measured by multiple observations and mother report. *Child Development*, **65**, 1478–1490.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, **7**, 301–317.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420–428.

Spinath, F. M., Angleitner, A., Borkenau, P., Riemann, R., & Wolf, H. (2002). German Observational Study of Adult Twins (GOSAT): A multimodal investigation of personality, temperament and cognitive ability. *Twin Research and Human Genetics*, **5**, 372–375.

Stifter, C. A., Willoughby, M. T., & Towe-Goodman, N. (2008). Agree or agree to disagree? Assessing the convergence between parents and observers on infant temperament. *Infant and Child Development*, **17**, 407–426.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing* (pp. 261–292). Thousand Oaks, CA: Sage.

Watson, D., & Clark, L. A. (1997). Extraversion and its positive emotional core. In *Handbook of personality psychology* (pp. 767–793). San Diego, CA: Academic Press.

Zahn-Waxler, C., Schmitz, S., Fulker, D., Robinson, J., & Emde, R. (1996). Behavior problems in 5-year-old monozygotic and dizygotic twins: Genetic and environmental influences, patterns of regulation, and internalization of control. *Development and Psychopathology*, **8**, 103–122.