

## SUPPLEMENT 1

## METHOD

## Pre-existing Individual Factors

*Birth Weight.* Each twin's birth weight was obtained with parental recall when the twins were 1 year old. Although parental recall is less accurate than obtaining birth weight directly from hospital records, a recent study reported that 85% of parents of 12- to 15-year-old children correctly recalled their children's birth weight to within  $\pm 227$  g.<sup>1</sup> In this substudy sample, birth weight ranged from 1135.0 to 3546.8 g (mean [SD] = 2292.6 [473.5]).

*IQ.* To assess children's IQ, each child was individually tested at age 5 years using a short form of the Wechsler Preschool and Primary Scale of Intelligence-Revised<sup>2</sup> comprising the Vocabulary and Block Design subtests. IQs scores were prorated following procedures described by Sattler.<sup>3</sup> In this substudy sample, IQ ranged from 76.0 to 135.0 (mean [SD] = 104.3 [13.7]).

*Internalizing and Externalizing Problems.* Internalizing and externalizing problems were assessed when the children were 5 years of age using the Child Behavior Checklist for mothers<sup>4</sup> and the Teacher's Report Form.<sup>5</sup> Mothers were given the instrument as a face-to-face interview, and teachers responded by mail. Both informants rated each item as being "not true," "somewhat or sometimes true," or "very true or often true." The reporting period was 6 months before the interview. The Internalizing problems scale is the sum of items on the Withdrawn and Anxious/Depressed subscales, including items such as "cries a lot," "withdrawn," "does not get involved with others," and "worries." The internal consistency reliabilities of the mother and teacher at 5 years were 0.83 and 0.85, respectively. The Externalizing problems scale is the sum of items from the Delinquency and Aggression scales, including items such as "gets in many fights," "lying or cheating," and "screams a lot." The internal consistency reliabilities of the mother and teacher at 5 years were 0.89 and 0.93, respectively. The internal consistency reliability of the combined mothers' and teachers' scales for internalizing and externalizing problems were 0.85 and 0.92, respectively. In this substudy sample, internalizing problems ranged from 0 to 40 (mean [SD] = 9.4 [7.1]) and from 1 to 51 (mean [SD] = 17.3 [11.6]) for externalizing problems.

## Child-Specific Family Environments

*Lifetime Maltreatment.* At each home visit, mothers were interviewed about past and ongoing physical harm that had happened to her children using the standardized clinical interview protocol from the Multi-Site Child Development Project, which has established validity and reliability in this sample and others.<sup>6</sup> Interviewers coded the likelihood that children had been physically maltreated based on the mother's narratives. Two independent raters (a senior investigator with clinical psychology background and a project leader with experience in social service) reviewed the notes for all families where maltreatment was reported to confirm the interviewers' ratings. Examples of maltreatment in our sample included: the mother smacked the child weekly and left marks or bruises; child was repeatedly beaten by a young adult stepsibling; children were routinely smacked by father when drunk; child was fondled sexually and slapped by their mother's boyfriend; children were frequently hit in the face, etc. Many, but not all, cases in the course of our research were under investigation by police or social services, with children already on the child-protection register or in foster care at follow-up, having been removed from their parents because of abuse. The research team followed the guidelines for referral of families under the United Kingdom's Children Act.<sup>7</sup> In this substudy sample of MZ twins, maltreatment has been documented in six families (20%) where both bullied and nonbullied twins were targeted.

*Maternal Warmth.* We assessed maternal warmth<sup>8</sup> using procedures adapted from the Five Minute Speech Sample method.<sup>9</sup> Mothers were asked to speak for 5 minutes about each of their children when they were 5 years of age. Warmth is a global measure of the whole speech sample and was assessed by the tone of voice, spontaneity, sympathy, and empathy toward the child. Warmth was coded on a 6-point scale. "High warmth" and "moderately high warmth" were coded when there was definite warmth, enthusiasm, interest in and enjoyment of the child. "Moderate warmth" was coded when there was definite understanding, sympathy and concern but only limited warmth of tone. "Some warmth" was coded when there was a detached and rather clinical approach, with little or no warmth of tone, but moderate understanding, sympathy and concern. "Very little warmth" was rated when there was only a slight amount of under-

standing, sympathy, or concern or enthusiasm about or interest in the child. "No warmth" was reserved for respondents who showed a complete absence of warmth. Two trained raters coded the tapes of the mothers' speech sample. Interrater reliability was established by having the raters individually code audio-tapes describing 40 children. The interrater agreement for maternal warmth was  $r = 0.90$ . In this substudy sample, maternal warmth ranged from 1 to 5 (mean [SD] = 3.4 [1.1]).

*Lifetime Stressful Life Events.* We assessed lifetime stressful life events by asking each twin if they ever experienced a series of age-relevant life events. We selected a total of 41 items from the Coddington Life Events Questionnaire (CLEQ)<sup>10,11</sup> and from the Traumatic Events Screening Inventory adapted for children.<sup>12</sup> Items included events such as major accidents, birth of a sibling, moving to a new school, hospitalization, witnessing domestic violence, disasters, and abuses. The Lifetime Stressful Life Events scale corresponds to the sum of the total number of events reported by each child, which ranged, in this substudy sample, from 2 to 24 (mean [SD] = 12.3 [4.8]).

#### Concomitant Individual Factors

*Body Mass Index.* Body mass index was calculated as children's weight in kilograms divided by the square of their height in meters, as measured on the day of the laboratory visit. In this substudy sample, body mass index ranged from 14.6 to 29.2 (mean [SD] = 20.1 [3.6]).

*Pubertal Stage.* Puberty stage was assessed when the twins were aged 12 using the Physical Development Scale,<sup>13</sup> a self-report measurement of both general (growth spurt in height, body hair, skin change) and gender-specific characteristics of pubertal development (boys: facial hair growth and voice change; girls: breast development and menarche). For each item, children responded whether there had been "no development" (0), "development had already begun" (1), "development was definitely underway" (2), or "development was already completed" (3). Menarche was coded dichotomously as menstrual periods had started (3) or had not (0). A total score was computed by summing scores on the five items. In this substudy sample, pubertal stage ranged from 0 to 16 (mean [SD] = 7.7 [3.9]).

*Bullying Perpetration.* As part of the age-12 assessments of children's disruptive behavior, we asked mothers and teachers whether children

had been bullying others. Mother reported that a total of 15.0% of children were bullies ( $n = 9$ ), whereas a similar statement was made about 18.9% of children by teachers ( $n = 10$ ; seven twins with missing data for the teacher's assessment). A child was considered to be a bully if reported by either source. A total of 18 children (30.0%) bullied others according to mothers and/or teachers.

*Family Socioeconomic Status.* Family socioeconomic status was defined through a standardized composite of parents' income, education and social class when children were 5 years of age.<sup>14</sup> The three SES indicators were highly correlated (correlations ranged from 0.57 to 0.67,  $P < .05$ ) and loaded significantly onto one latent factor (factor loadings = 0.80, 0.70, and 0.83 for income, education, and social class, respectively). The population-wide distribution of the resulting factor was divided in tertiles. In this substudy sample, one-fourth of the children were living in a low SES situation (26.7%).

#### Stress-Related Measures

*Perceived Stress.* We assessed twins' levels of perceived stress after the PST using an adaptation of the Perceived Stress Scale<sup>15</sup> to assess whether they experienced the laboratory challenge as unpredictable, uncontrollable and unmanageable. The Perceived Stress Scale included eight items enquiring whether, during the challenge children felt that "what I was asked to do was just too much" or "I could not cope with what I had to do." Children rated each item as being "very slightly or not at all" to "extremely" true. In this substudy sample, perceived stress ranged from 0 to 22 (mean [SD] = 8.1 [4.8]).

*Negative Affective Scale.* To evaluate whether the laboratory challenge elicited negative emotions, twins were asked to fill out the Positive and Negative Affective Scales before and immediately after the PST.<sup>16</sup> This instrument is composed of 20 items assessing positive and negative mood. Only the confounding effect of the negative scale was examined. Twins completed the negative scale before and after the PST, rating emotions specific to negative moods (e.g., ashamed, upset, and distressed) as being "very slightly or not at all" to "extremely" true. The increase of negative affect was calculated by subtracting the sum of the 10 items assessed before the PST from those evaluated after. In this

substudy sample, the increase in negative affect ranged from  $-5$  to  $16$  (mean [SD] =  $2.0$  [ $4.1$ ]).

## REFERENCES

- Walton KA, Murray LJ, Gallagher AM, Cran GW, Savage MJ, Boreham C. Parental recall of birthweight: a good proxy for recorded birthweight? *Eur J Epidemiol.* 2000;16:793-796.
- Wechsler D. Wechsler Preschool and Primary Scale of Intelligence—Revised. London: Psychological Corporation, Harcourt Brace and Company; 1990.
- Sattler J. Assessment of Children: WISC-III and WPPSI-R Supplement. San Diego: Sattler; 1992.
- Achenbach TM. Manual for the Child Behavior Checklist 4-18 and Profile. Burlington, VT: University of Vermont, Department of Psychiatry; 1991.
- Achenbach TM. Manual for the teacher's report form and 1991 profile. Burlington, VT: University of Vermont, Department of Psychiatry; 1991.
- Dodge KA, Bates JE, Pettit GS. Mechanisms in the cycle of violence. *Science.* 1990;250:1678-1683.
- Department of Health. The Children Act. London: Her Majesty's Stationery Office; 1989.
- Caspi A, Moffitt TE, Morgan J, *et al.* Maternal expressed emotion predicts children's antisocial behavior problems: using monozygotic-twin differences to identify environmental effects on behavioral development. *Dev Psychol.* 2004;40:149-161.
- Magana AB, Goldstein JM, Karno M, Miklowitz DJ, Jenkins J, Falloon IR. A brief method for assessing expressed emotion in relatives of psychiatric patients. *Psychiatry Res.* 1986;17:203-212.
- Coddington RD. The significance of life events as etiologic factors in the diseases of children. I. A survey of professional workers. *J Psychosom Res.* 1972;16:7-18.
- Coddington RD. The significance of life events as etiologic factors in the diseases of children. II. A study of a normal population. *J Psychosom Res.* 1972;16:205-213.
- Ribbe D. Psychometric review of Traumatic Event Screening Instrument for Children (TESI-C). In: Stamm BH, ed. Measurement of Stress, Trauma, and Adaptation. Lutherville, MD: Sidran Press; 1996:386-387.
- Petersen AC, Crockett L, Richards M, Boxer A. A self-report measure of pubertal status: reliability, validity, and initial norms. *J Youth Adolesc.* 1988;17:117-133.
- Trzesniewski KH, Moffitt TE, Caspi A, Taylor A, Maughan B. Revisiting the association between reading achievement and antisocial behavior: new evidence of an environmental explanation from a twin study. *Child Dev.* 2006;77:72-88.
- Cohen S, Williamson G. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, eds. *The Social Psychology of Health.* Newbury Park, CA: Sage; 1988:31-67.
- Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol.* 1988;54:1063-1070.