Journal of Abnormal Child Psychology, Vol. 33, No. 5, October 2005, pp. 537–550 (© 2005)
DOI: 10.1007/s10802-005-6736-5

Psychometric Evaluation of 5- and 7-Year-Old Children's Self-Reports of Conduct Problems

Louise Arseneault,[1,3] Julia Kim-Cohen,[1,2] Alan Taylor,[1] Avshalom Caspi,[1,2] and Terrie E. Moffitt[1,2]

Received November 18, 2003; revision received November 9, 2004; accepted December 10, 2004

Past research suggests that young children are incapable of reporting information about their own behavior problems. To test this, we examined the validity and the usefulness of children's self-reports in the E-Risk Study, a nationally representative birth cohort of 2,232 children. We used the Berkeley Puppet Interview to obtain children's self-reports of conduct problems when they were 5-years old and the Dominic-R when they were 7-years old. We also collected information about the children and their families by interviewing mothers, sending questionnaires to teachers, and rating examiners' observations during home visits. Results indicate that when children's self-reports are gathered with structured and developmentally appropriate instruments, they are shown to be valid measures: conduct problems reported by the children themselves were associated with known correlates including individual characteristics (e.g., IQ), related behaviors (e.g., hyperactivity), and family variables (e.g., economic disadvantages). Observed correlations closely matched effect sizes reported in the literature using adults' reports of children's behavioral problems. In addition, children's self-reports can be useful: both measures distinguished children meeting *DSM-IV* criteria for research diagnoses of conduct disorder. Children's reports also contributed unique information not provided by adults. For research and clinical purposes, young children's self-reports can be viewed as a valuable complement to adults' ratings and observational measures of children's behavior problems.

KEY WORDS: children self-reports; conduct problems; antisocial behavior; validity.

Children are generally thought to be unable to accurately report about their own disruptive behavior (Boyle et al., 1993; Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985; Schwab-Stone, Fallon, Briggs, & Crowther, 1994). The aim of this study is to report data on the validity and the usefulness of young children's self-reports of conduct problems.

Children's self-reports can offer a valuable complement to more traditional methods of obtaining information about children's behavior. Indeed, children's self-reports of socially undesirable behaviors (e.g., stealing, swearing) may be a more accurate measure compared to adults' reports because children usually conceal such acts from their parents or teachers and because children have an unrestricted awareness of their own behavior across settings. Given the limits of adult reports of children's conduct problems, children themselves could provide valuable information about their own activities and behaviors, but can young children validly report these phenomena? There are several challenges to obtaining reliable and valid data from children about their own behavior (La Greca, 1990; Stone & Lemanek, 1990). For example, children's vocabulary and cognitive limitations may reduce their understanding of some interview questions. Language difficulties can also impede the complete disclosure of answers. As a consequence, the interview might become a frustrating experience for the children and a difficult task for the interviewer. In view of children's short attention span, responding to long questionnaires may translate into a tedious and

[1]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, London.
[2]Department of Psychology, University of Wisconsin, Madison, Wisconsin.
[3]Address all correspondence to Louise Arseneault, SGDP Centre, Institute of Psychiatry, Box Number P080, De Crespigny Park, London SE5 8AF, United Kingdom; e-mail: l.arseneault@iop.kcl.ac.uk.

boring task. Attention difficulties may also give rise to behavioral problems during the assessment. These potential problems raise concerns about young children's abilities to participate in interview formats typical of adult assessments and may explain why children have been considered as unreliable reporters about their own behavior. However, recent years have witnessed an increased recognition that children's self-reports would be valuable tools if new methodologies were developed to overcome children's limited abilities (see Warren, Oppenheim, & Emde, 1996).

Studies have shown the feasibility of collecting self-reported measures of disruptive behavior from young children using two developmentally appropriate instruments designed to assess mental health symptomatology among young children (Ablow et al., 1999; Valla, Bergeron, Bidaut-Russell, St-Georges, & Gaudet, 1997; Valla, Bergeron, & Smolla, 2000). Firstly, the Berkeley Puppet Interview (BPI) was designed to assess nine mental health components (e.g., depression, inattention, conduct problems) among 4- to 8-year-old children (Measelle, Ablow, Cowan, & Cowan, 1998). The BPI uses puppets to engage young children in a friendly, conversational interview in which children are invited to talk about themselves. Secondly, the Dominic-R (Valla, Bergeron, Bérubé, Gaudet, & St-Georges, 1994) is a structured interview that was developed to assess symptoms of the seven most prevalent *DSM-III-R* Axis-I mental disorder diagnoses (e.g., anxiety, oppositional defiant disorders, conduct problems) in children aged 6–11 years. The Dominic-R uses drawings to capture children's interest and to illustrate behaviors targeted by the interview. Although past research supports the use of these two age-appropriate instruments, the validity of children's self-reports of conduct problems has not yet been extensively evaluated in a large representative sample.

The validity of any given measure is established by examining the pattern of associations between the measure and a set of theoretically related constructs (Flanery, 1990; Westen & Rosenthal, 2003). Current theories and a large body of supportive empirical findings indicate several correlates and features of children's conduct problems (Lahey, Moffitt, & Caspi, 2003). In general, children's disruptive behaviors are positively associated with risk indicators such as socioeconomic disadvantage, harsh discipline, and other behavioral problems. Disruptive behaviors are also negatively associated with resilience factors such as high IQ. The overall magnitude of the associations between adults' ratings of children's conduct problems and their correlates are, in general, moderate. If young children's self-reports of conduct problems are valid, we should observe similar patterns of associations.

In addition to being valid, a measure has to be useful—that is, provide valuable and unique information for clinical and research purposes. For clinicians assessing children's symptomatology, a useful measure of young children's conduct problems should distinguish children with and without a diagnosis of conduct disorder (Keenan & Wakschlag, 2002; Kim-Cohen et al., 2005). For researchers gathering information mainly from adults, a useful measure of children's conduct problems should carry unique information not reported by any other informants.

This article reports on two studies of the validity and the usefulness of young children's self-reports of conduct problems. In Study 1, we examined self-reports of conduct problems when children were 5-years old using the Berkeley Puppet Interview, and in Study 2, when children were 7-years old using the Dominic-R. We examined the validity of each measure by reporting associations between children's self-reported conduct problems and several correlates. We examined the usefulness of each measure by comparing children with and without a research diagnosis of conduct disorder on the two self-report measures and by testing the unique additional contribution of children's self-reports to the prediction of adults' reports of children's conduct problems, over and above information provided by other adult raters.

METHOD

The E-Risk Study Sample

Participants are members of the Environmental Risk (E-Risk) Longitudinal Twin Study, which investigates how genetic and environmental factors shape children's development. The study follows an epidemiological sample of families with young twins who were interviewed in the home when the twins were age 5 and 7 years. The E-Risk sampling frame was two consecutive birth cohorts (1994 and 1995) in a birth register of twins born in England and Wales (Trouton, Spinath, & Plomin, 2002). Of the 15,906 twin pairs born in these 2 years, 71% joined the register.

The E-Risk Study sought a sample size of 1,100 families to allow for attrition in future years of the longitudinal study while retaining statistical power. An initial list of families who had same-sex twins was drawn from the register to target for home visits, with a 10% oversample to allow for nonparticipation. The probability sample was drawn using a high-risk stratification sampling frame. High risk families were those in which the mother had her first birth when she was 20 years of age or younger. We used this sampling (1) to replace high risk families who

were selectively lost to the register via nonresponse and (2) to ensure sufficient base rates of problem behavior given the low base rates expected for 5-year-old children. Age at first childbearing was used as the risk-stratification variable because it was recorded for virtually all families in the register, it is relatively free of measurement error, and early childbearing is a known risk factor for children's problem behaviors (Maynard, 1997; Moffitt & E-Risk Study Team, 2002). The sampling strategy resulted in a final sample in which two-thirds of Study mothers accurately represent all mothers in the general population (aged 15–48) in England and Wales in 1994–95 (estimates derived from the General Household Survey; Bennett, Jarvis, Rowlands, Singleton, & Haselden, 1996). The other one-third of Study mothers (younger only) constitute a 160% oversample of mothers who were at high risk based on their young age at first birth (15–20 years). To provide unbiased statistical estimates that can be generalized to the population of British families with children born in the 1990s, the data reported in this article were corrected with weighting to represent the proportion of young mothers in that population.

Of the 1,203 families from the initial list who were eligible for inclusion, 1,116 (93%) participated in home-visit assessments when the twins were age 5 years forming the base sample for the study: 4% of families refused, and 3% were lost to tracing or could not be reached after many attempts. In the sample overall, 90.6% of twin pairs were Caucasian, 4.1% were Asian, 1.4% were Black, and 3.9% were mixed race or "other;" 82% of the mothers were currently living with the biological fathers of the twins. With parent's permission, questionnaires were posted to the children's teachers, and teachers returned questionnaires for 94% of cohort children. After complete description of the study to the participants, written informed consent was obtained from mothers. This first visit will be referred to as the age-5 assessment.

A follow-up home visit was conducted 18 months after the age-5 assessment when the twins were 6.5-years old on average (range 6.0–7.0 years). Follow-up data were collected for 98% of the 1,116 E-Risk Study families. At this follow-up, teacher questionnaires were obtained for 91% of the 2,232 E-Risk Study twins (93% of those taking part in the follow-up). This follow-up will be referred to as the age-7 assessment.

Measures

Self-Reports of Conduct Problems at Ages 5 and 7 Years

We used the Berkeley Puppet Interview (BPI; Measelle et al., 1998) to obtain self-reports from the children about their own disruptive behavior at age 5 years. In the BPI, the examiner introduces two identical fluffy animal puppets (Iggy and Ziggy) to the child, and the puppets invite the child to join them in a conversation in which they tell the child things about themselves and the child tells them about him/herself. The two puppets make opposite statements (e.g., Iggy: "I hit kids a lot"—Ziggy: "I don't hit kids") in a counterbalanced order. The puppets then ask the child to tell how he/she behaves. Children are allowed to indicate their answer verbally, or nonverbally by pointing or touching the puppet. The BPI was administered to each twin separately. Interviews were videotaped to score the children's answers later. All examiners completed a 1-week certification-training course designed by Ablow and Measelle (1999).

The children were administered 19 items covering three BPI scales that assess disruptive behavior (items are listed on Table II): Overt Aggression/Hostility, Conduct Problems, and Oppositionality. Two different coders scored each interview, with interrater reliability exceeding .90 for all coders. Every item was coded on a Likert scale ranging from 1 (*no symptom*) to 7 (*definite symptom*). Scores at both extremes of the scale were given when children amplified their answers or used superlatives (e.g., "I never hit kids"). When children endorsed the puppet's statements (e.g., "I don't hit kids") or provided nonverbal responses, coders rated these answers as less extreme with scores of 2 or 6. If the child modified his/her answer or added a condition (e.g., "I don't hit kids at school"), scores of 3 or 5 were given by the coder. A score of 4 represented rare cases where children agreed with both puppets. Scores ranged from 31 to 106 ($M = 51.45$, $SD = 13.35$, Median = 46.94), and the internal consistency reliability was .81. The test–retest reliability for the three subscales ranged from .52 to .69 in clinical and community samples (Ablow et al., 1999). Data for the BPI were missing for 353 children, leaving valid data for 84% of the sample children. Missing data was mainly caused by procedural effects (e.g., lack of time or lack of privacy for the interview). In a few instances, the child could not complete the interview or the examiner and/or the coder assessed that the child did not understand the task.

The use of puppets was not age-appropriate for a second behavioral assessment with our sample of 7-year-olds. Therefore, we used the Dominic-R (Valla et al., 1997), an interview using visual and auditory stimuli, to collect self-reports of conduct problems during the age-7 assessment. The interviewer first presents the child with a booklet containing drawings (visual stimuli) depicting Dominic, a gender-neutral character, in various situations. Each drawing is accompanied by a question about its specific behavioral content. The interviewer reads aloud

questions to the child (auditory stimuli) (e.g., "Have you ever hurt people on purpose like Dominic?"). Questions are printed beneath the drawings so both the interviewer and the child can read them simultaneously. The child is then asked to tell whether he/she has behaved like Dominic in the past. The child's answers were scored (0 = *no*; 1 = *yes*) on a separate sheet immediately after each item. The interview was administered to each twin individually.

We asked nine items covering the Dominic-R's Conduct Disorder scale including items on physical aggression (items are listed on Table IV). The majority of children did not report any conduct problems at age 7 years ($N = 1,514$, unweighted; 78.6%), 14.6% answered positively to one item ($N = 300$), and the remainder reported two or more disruptive behaviors ($N = 147$, 6.8%). The test–retest reliability has already been reported to be .71 in a group of 7-year-olds (Valla et al., 1997). Data for the Dominic-R were missing for 271 children at age 7, leaving valid data for 88% of the sample children. A high proportion of missing data with the Dominic-R was due to procedural effects.

Mothers' and Teachers' Ratings of Conduct Problems at Ages 5 and 7 Years

Mothers' and teachers' reports of children's disruptive behavior at ages 5 and 7 years were collected using the Achenbach family of instruments, namely the Child Behavior Checklist (CBCL; Achenbach, 1991a) for the mothers, and the Teacher's Report Form (TRF; Achenbach, 1991b) for the teachers. Mothers were given the instrument as a face to face interview and teachers responded by post. Both informants rated each item as being *not true* (0), *somewhat* or *sometimes true* (1), or *very true* or *often true* (2). The reporting period was 6 months prior to the interview. Children's disruptive behavior was assessed with 43 items from the Delinquency and Aggression scales, supplemented with items from the Diagnostic and Statistical Manual of Mental Disorders (*DSM-IV*; American Psychiatric Association, 1994) assessing conduct (e.g., "uses force to take something from another child") and oppositional defiant disorder (e.g., "spiteful, tries to get revenge"). At age 5 years, mothers' scores ranged from 0 to 72 ($M = 15.52$, $SD = 11.41$) and teachers' scores ranged from 0 to 74 ($M = 5.65$, $SD = 9.09$). The internal consistency reliabilities for disruptive behavior were .92 for the mothers' reports and .94 for the teachers' reports. At age 7 years, mothers' scores ranged from 0 to 72 ($M = 13.18$, $SD = 10.62$) and teachers' scores ranged from 0 to 66 ($M = 5.29$, $SD = 8.68$). The internal consistency reliabilities were .93 for the mothers' reports and .95 for the teachers' reports.

To assess children's behavioral problems in the clinical range, we derived a research diagnosis of children's conduct disorder on the basis of mothers' and teachers' reports on 14 of 15 *DSM-IV* symptoms of conduct disorder. The "forced sexual activity" symptom was excluded as inappropriate for 5-year-olds. A child was considered to have a given symptom if either the mother or the teacher reported the symptom as being "very true or often true" (score = 2) in the past 6 months. We counted a symptom as present if there was evidence of it from either source, following evidence that this approach enhances diagnostic validity (Bird, Gould, & Staghezza, 1992; Piacentini, Cohen, & Cohen, 1992). The most frequently endorsed symptoms were "deliberately destroys others' property," "starts fights," and "uses force to take things from others." Consistent with *DSM-IV* criteria, children with three or more symptoms were assigned a research diagnosis of conduct disorder. The prevalence of children with a research diagnosis of conduct disorder in the sample was 6.6% ($N = 189$, unweighted). By age 7, 234 children (8.0%) met *DSM-IV* criteria for a research diagnosis of conduct disorder.

Examiner-Observers' Ratings of Conduct Problems at Age 5 Years

After the home visit, interviewers rated each child on the Dunedin Behavioural Observation Scale, which includes 9 items measuring disruptive behavior (e.g., hostility, lability, roughness; Caspi, Henry, McGee, Moffitt, & Silva, 1995). Each behavior was defined in explicit terms, and the interviewer evaluated whether each characteristic was observed (0) *not at all*, (1) *somewhat*, or (2) *definitely during the home visit*. Scores ranged from 0 to 18 ($M = 2.22$, $SD = 3.46$). The internal consistency reliability of the examiner report of disruptive behavior was .90 and the inter-rater reliability coefficient was .70.

Behavioral Observations at Age 5 Years

The *Snap*! is a rigged competitive card game that allows for direct observations of children's disruptive behavior in a potentially threatening situation, i.e., losing to another child. The game involves matching pictures on cards and was adapted from an instrument developed by Murray and colleagues (Murray, Woolgar, Cooper, & Hipwell, 2001). The cards are rigged so that each child is exposed to a winning and a losing streak in counterbalanced order. On the final deal, both children emerge as joint winners. The game was videotaped during the home-visits and a trained researcher coded the child's

behavior on the left-hand side before returning to code the cotwin on the right-hand side of the screen. Acts of disruption included cheating, knocking the board over, throwing the counters, swearing or other forms of verbal aggression, hitting the playmate, and storming out of the room. Ratings ranged from 1 (*child cooperative throughout the game*) to 5 (*child's disruptive behavior results in premature game termination*). The interrater reliability coefficient was .83 (Hughes et al., 2002).

Measures of Children's Internalizing Problems, Hyperactivity Problems, and Prosocial Behavior at Ages 5 and 7 Years

Internalizing problems were assessed using the Child Behavior Checklist (CBCL; Achenbach, 1991a) for the mothers, and the Teacher's Report Form (TRF; Achenbach, 1991b) for the teachers. The internalizing problems total scale is the sum of items in the Withdrawn, Somatic Complaints, and Anxious/Depressed scales including items such as "cries a lot," "feels too guilty," and "worries." At age 5 years, mothers' scores ranged from 0 to 44 ($M = 8.35$, $SD = 6.68$). Teachers' scores ranged from 0 to 50 ($M = 5.85$, $SD = 5.76$). The internal consistency reliabilities of the mothers' and teachers' reports were .84 (31 items) and .85 (35 items), respectively. At age 7 years, mothers' ratings ranged from 0 to 43 ($M = 7.32$, $SD = 6.21$) and teachers' ratings ranged from 0 to 46 ($M = 5.79$, $SD = 6.01$). The internal consistency reliabilities were .86 for mothers' reports and .87 for teachers' reports.

Children's hyperactivity was measured with 18 items from the Rutter Child Scales (Sclare, 1997) and supplemented with items concerning inattention, impulsivity, and hyperactivity derived from the *DSM-IV* diagnostic criteria for Attention Deficit Disorder (e.g., "cannot settle to anything for more than a few moments, quickly moves from one thing to another," "fidgety or squirmy"). At age 5 years, mothers' and teachers' ratings ranged from 0 to 34 (mothers: $M = 10.38$, $SD = 7.49$; teachers: $M = 5.02$, $SD = 6.55$). The internal consistency reliabilities of the mothers' and teachers' reports were .90 and .94, respectively. At age 7 years, scores varied from 0 to 34 for both mothers and teachers (mothers: $M = 9.27$, $SD = 7.22$; teachers: $M = 4.43$, $SD = 6.32$). The internal consistency reliabilities were .91 for mothers' reports and .94 for teachers' reports.

Prosocial behavior was measured with 10 items from the Revised Rutter Scale for School-Age Children (Goodman, 1994; Sclare, 1997), including items such as "tries to be fair in games," and "considerate of other people's feelings." At age 5 years, ratings from both mothers and teachers ranged from 0 to 20 (mothers: $M = 16.31$, $SD = 3.28$; teachers: $M = 11.74$, $SD = 4.86$). The internal consistency reliabilities of parents' and teachers' reports were .76 and .92, respectively. At age 7 years, scores varied from 0 to 20 according to mothers' ratings ($M = 16.40$, $SD = 3.32$), and also teachers' ratings ($M = 12.71$, $SD = 4.80$). The internal consistency reliabilities were .80 for mothers' reports and .93 for teachers' reports.

Cognitive Ability and Achievement

At age 5 years, children's IQ was assessed with a short form of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R; Wechsler, 1990). Using two subtests (Vocabulary and Block Design), children's IQs were computed following procedures described by Sattler (1992). IQ scores ranged from 52 to 145 and the sample mean was 97.83 ($SD = 14.40$).

At age 7 years, questions about children's academic achievement were included in the TRF (Achenbach, 1991b). Teachers were asked whether the child's current mathematical and English performances were (1) far below average; (2) somewhat below average; (3) average; (4) somewhat above average; or (5) far above average, compared to pupils of the same age. Scores were averaged across topics to give a global scale of school performance. Scores ranged from 1 to 5 ($M = 3.02$, $SD = .93$).

At age 7 years, children's reading abilities were individually tested using the Test Of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). The TOWRE provides a quick assessment of sight word efficiency. The sight word efficiency measures the number of real printed words that can be accurately identified in 45's and provides an index of the size of the child's reading vocabulary. The children's scores were converted to age-based standard scores (with a score of 100 = median). The children in this study had an average sight word efficiency score of 105.90 ($SD = 12.96$).

Socioeconomic Disadvantage

The socioeconomic disadvantage scale is a count of seven socioeconomic disadvantages, which were defined as follows: (1) head of household has no educational qualifications; (2) head of household is employed in an unskilled occupation or is not in the labor force; (3) total household gross annual income is less than £10,000; (4) family receives at least one government benefit, excluding disability benefit; (5) family housing

is government subsidized; (6) family has no access to a vehicle; and (7) family lives in the poorest of six neighborhood categories, in an area dominated by government-subsidized housing, low incomes, high unemployment, and single-parent families. Summing across these seven items yielded a composite index of socioeconomic disadvantage, ranging from 0 to 7 ($M = 1.19$, $SD = 1.71$).

Parent's Antisocial Behavior

We interviewed mothers about their own histories of antisocial behavior using the Young Adult Self Report (YASR; Achenbach, 1997), modified to obtain lifetime data. We report scores on the externalizing syndrome, which is the sum of 39 items on the scales of Delinquent Behavior and Aggressive Behavior. Mothers rated each behavior as being *not true* (0), *somewhat* or *sometimes true* (1), or *very true* or *often true* (2). Scores ranged from 0 to 60 ($M = 11.25$, $SD = 9.71$) and the internal consistency reliability of this scale was .90. Mothers also reported about the biological fathers' lifetime histories of antisocial behavior using the same instrument. Scores for biological fathers' antisocial behavior ranged from 0 to 88 ($M = 14.76$, $SD = 16.29$) and the internal consistency reliability of this scale was .95 (44 items). A methodological study of mother–father agreement attests to the reliability of these women's reports about men's problem behaviors; mothers' reports account for more than 75% of the variance in men's self-reports on these scales (Caspi et al., 2001).

Parenting Quality

Corporal punishment was assessed separately for each twin when they were age 5 years by interviewing mothers with the standardized clinical interview protocol from the Multi-Site Child Development Project (Dodge, Bates, & Pettit, 1990; Dodge, Pettit, Bates, & Valente, 1995; Landsford et al., 2002). Mothers were asked whether they had used a variety of disciplinary practices, some of which assessed corporal punishment: "grabbing or shaking," "smacking or hitting," or engaging in "other physical discipline." A score of 1 was assigned if the mother reported that she had used a particular disciplinary practice with her child and a score of 0 was assigned if she had not. If mothers reported that they used any form of corporal punishment, they were then asked how often in the past year the child was physically punished, with responses ranging from 0 (*never*) to 5 (*daily*). Mothers who did not engage in any kind of corporal punishment were assigned a score of 0 on the frequency variable. We created "variety" and "frequency" variables based on mothers' reports of corporal punishment across the children's first 5 years. The variety and frequency scores across the child's first 5 years were highly correlated ($r = .65$, $p \leq .001$). A corporal punishment composite variable was created by standardizing and summing the variety and frequency scores. Scores ranged from -3.29 to 7.33 ($M = .03$, $SD = 1.78$). 87% of children had experienced corporal punishment at least once in their first 5 years.

Maternal expressed emotion was assessed as part of a 5 min speech sample to elicit expressed emotion during the home visit at age 5 years. Trained interviewers asked the mother to describe each of their children ("For the next 5 min, I would like you to describe [child] to me, what is [child] like?"). The mother was encouraged to talk freely with few interruptions. For this study, we examined 2 variables: negativity (mothers making disparaging remarks and finding fault with the child; resentment and hostility towards the child) and warmth (definite and clear-cut tonal warmth, enthusiasm, interest in, and enjoyment of the child). All interviews were audiotaped with the mother's consent. Two trained raters coded the audiotapes according to developmentally appropriate guidelines for scoring expressed emotion with preschool children (Caspi et al., 2004). A six-point rating scale refers to the degree of negativism and warmth expressed in the interview about the child. Negativity scores ranged from 0 to 5 ($M = 1.46$, $SD = .93$) and warmth scores from 0 to 5 ($M = 3.36$, $SD = .98$). The interrater agreement rate was .84 for the negativity scale and .90 for the warmth scale.

On completion of the home visit when the children were age 7 years, interviewers completed a questionnaire asking about various aspects of the family's life such as the physical (e.g., safe, clean and conducive to health development), cognitive (e.g., growth-fostering materials), and emotional (e.g., chaos, affection) climate in the home. This questionnaire was based on the Home Observation for Measurement of the Environment (HOME; Caldwell & Bradley, 1984) and the University of Washington Parenting Clinic (Webster-Stratton, 1998). Interviewers were trained to observe family interactions and the quality of the home environment. The response format was a 3-point scale: *no* (0), *a little*/*somewhat* (1), *yes* (2). Chaos in the house was measured with three items such as "Is the house chaotic or overly noisy?" Scores ranged from 0 to 6 ($M = 1.09$, $SD = 1.41$). Negative parenting was measured with 7 items such as "Was the parent controlling?" and "Was the parenting erratic, inconsistent or haphazard?" Scores ranged from 0 to 14 ($M = .79$, $SD = 1.73$).

Neglect was measured with 6 items such as "Is the twin well nourished?" Scores ranged from 0 to 12 ($M = .73$, $SD = 1.51$). The internal consistency reliabilities for these scales were .53, .78, and .74 respectively. The interrater reliability coefficients were .86, .92, and .78 respectively.

Mothers' Experience of Domestic Violence

Adult domestic violence was assessed by inquiring about 12 acts of physical violence, including all nine items from the Conflict Tactics Scale Form R (Straus, 1990), plus three additional items describing other physically abusive behaviors such as "pushed/grabbed/shoved," "thrown bodily," and "threatened with knife/gun." Mothers were asked about their own violence toward any partner and about any partners' violence toward them during the last 5 years since the twins' birth, responding *not true* (0) or *true* (2). Another response option, *somewhat true* (1), was available for mothers who felt uncertain about their responses, but it was virtually unused by the mothers. Scores were summed (range = 0–40, $M = 2.76$, $SD = 5.67$). The internal consistency reliability of the physical abuse scale was .89. Additional methodological research shows that interpartner agreement for this measure is very high (latent correlation = 0.77; Moffitt et al., 1997).

Statistical Methods

We measured associations between children's self-reported conduct problems and their correlates by using Pearson correlations for BPI data. For Dominic-R data, we analyzed a series of planned comparisons using sets of contrast codes (Rosenthal & Rosnow, 1985). Statistical analyses of data were complicated by the fact that our twin study contained two children from each family, leading to nonindependent observations. As such, we analyzed data using standard regression techniques, but with tests based on the sandwich or Huber/White variance estimator (Rogers, 1993; Williams, 2000), a method available in STATA 8.0 (StataCorp, 2003). This technique adjusts estimated standard errors to account for the dependence in the data.

We evaluated group differences between children with and without a research diagnosis of conduct disorder on items from the self-reported instruments using *t*-tests (for BPI items ranging on a continuous scale) and odd ratios (for the dichotomous Dominic-R items). We calculated the effect sizes of the obtained group differences, using the formula:

$$d = (M_1 - M_0)/sd$$

where M_1 is the mean for the sample of children with a conduct disorder research diagnosis, and M_0 is the mean for the sample of children without a research diagnosis of conduct disorder, and *SD* is the standard deviation taken over the whole sample. For dichotomous variables, we estimated the standardized mean difference statistic (*d*) by taking the product of the log odds ratio and (sqrt)$3/p$ (Haddock, Rinkdskopf, & Shadish, 1998).

We assessed the additional value of children's reports about their own behavior, for Study 1, by conducting longitudinal regression analyses predicting teachers' reports of children's conduct problems at age 7 with BPI data collected at age 5 years, over and above mothers' and teachers' reports of children's conduct problems at age 5. We chose teachers' reports as the outcome measure, as opposed to mothers' reports, because teachers changed across assessments and this eliminates the possibility that shared-method variance underestimates the contribution of other informants. For Study 2, we conducted two separate regression analyses with concurrent measures, one predicting teachers' reports at age 7 with Dominic-R data collected at age 7 years, over and above mothers' reports at age 7, and one predicting mothers' reports at age 7 with self-reports over and above teachers' reports at age 7.

RESULTS

Study 1—The Berkeley Puppet Interview at Age-5 Years

Children's self-reported conduct problems were associated in the expected directions with known correlates of disruptive behavior (Table I). Children's self-reports were significantly associated with ratings from three different adult informants of children's conduct problems, and also an observational measure of children's behavior. BPI scores were correlated with gender, IQ, and social disadvantage. Children's self-reports of conduct problems did not correlate with either mothers' or teachers' reports of internalizing problems, but they did correlate with mothers' and teachers' ratings of hyperactivity and prosocial behavior. Children's reports of their own disruptive behavior were also correlated with mothers' and fathers' antisocial behavior, parenting quality, and domestic violence assessed at age 5 years.

Children with a research diagnosis of conduct disorder were more likely than nondiagnosed children to

Table I. Validity Analysis of the Berkeley Puppet Interview Assessing Conduct Problems at Age 5 Years

Correlates	Age-5 children's self-report	
	r	(95% CI)
Independent informants about conduct problems at age 5		
Mothers ($N = 1,877$)	.19**	(.15, .23)
Teachers ($N = 1,763$)	.21**	(.17, .26)
Examiner-observers ($N = 1,877$)	.20**	(.16, .25)
Behavioral observations ($N = 1,811$)	.16**	(.12, .21)
Gender		
0 = females; 1 = males	.14**	(.10, .19)
Cognitive abilities		
Age-5 IQ ($N = 1,875$)	−.25**	(−.29, − .20)
Family environment		
Age-5 social disadvantage ($N = 1,879$)	.17**	(.13, .22)
Age-5 internalizing problems		
Mothers' report ($N = 1,877$)	.04	(−.01, .08)
Teachers' report ($N = 1,759$)	−.04	(−.08, .01)
Age-5 hyperactivity problems		
Mothers' report ($N = 1,877$)	.18**	(.13, .22)
Teachers' report ($N = 1,763$)	.23**	(.18, .27)
Age-5 prosocial behavior		
Mothers' report ($N = 1,877$)	−.12**	(−.16, − .07)
Teachers' report ($N = 1,712$)	−.17**	(−.22, − .13)
Parents' antisocial behavior		
Mothers' antisocial behavior ($N = 1,873$)	.08**	(.03, .12)
Fathers' antisocial behavior ($N = 1,865$)	.10**	(.05, .14)
Parenting quality		
Age-5 corporal discipline ($N = 1,862$)	.06*	(.01, .10)
Age-5 EE negativism ($N = 1,684$)	.14**	(.10, .19)
Age-5 EE warmth ($N = 1,686$)	−.17**	(−.22, − .13)
Mothers' experience of domestic violence		
Age-5 domestic violence ($N = 1,879$)	.06∗	(.01, .10)

*$p < .05$. **$p < .001$.

endorse nearly all items of the BPI (Table II). There were only three exceptions: "I lose my temper," "I don't do what my mummy/daddy ask me to do," "I cheat when playing a game." However, given the overall pattern of results, the mean score on the BPI total scale was significantly higher for children with conduct disorder than children without conduct disorder. Effect sizes ranged from small to medium (Cohen, 1992).

Age-5 self-reported conduct problems predicted children's disruptive behavior two years later when the children were age 7 years according to both mothers' ($r = .17$, $N = 1839$, $p < .001$, 95% confidence intervals = .12, .21) and teachers' ratings ($r = .19$, $N = 1715$, $p < .001$, 95% confidence intervals = .14, .23). Moreover, the BPI uniquely contributed to the prediction of teachers' ratings of children's conduct problems at age 7 (standardized $\beta = .08$, $t = 3.06$, $p < .002$), over and above mothers' reports at age 5 ($\beta = .14$, $t = 4.72$, $p < .001$) and teachers' reports at age 5 ($\beta = .46$, $t = 11.10$, $p < .001$).

Study 2—The Dominic-R at Age-7 Years

Children's self-reports of conduct problems were consistent over a two year period despite the use of different instruments (Table III): mean scores on the BPI increased as positive endorsement of Dominic-R items increased from no symptom, to one symptom and to two symptoms. Agreement between informants was also found with age-7 data; both mothers' and teachers' ratings of disruptive behavior significantly increased in a dose-response fashion with higher numbers of conduct problems reported by the children themselves using the Dominic-R (Table III).

Associations with age-5 correlates of BPI conduct problems replicated when using the Dominic-R, and expected associations were also found between the Dominic-R and further measures collected during the age-7 assessment (Table III). Males were overrepresented among children who reported one or more conduct problems. Scores on the Dominic-R were

Table II. Mean Scores on Items From the Berkeley Puppet Interview (Coded on a Scale From 1 to 7) for Children With and Without a Research Diagnosis of Conduct Disorder (CD) at Age 5 Years

	Diagnosis				
	CD ($N = 146$)	No CD ($N = 1{,}733$)			
BPI items	*M* (SD)	*M* (SD)	*t*	(*df*)[a]	*d*
I lose my temper	3.5 (1.9)	3.6 (2.0)	0.71	(974)	0.05
I take things that don't belong to me	3.4 (1.8)	3.0 (1.7)	2.08	(970)*	0.23
It's funny when a kid gets in trouble at school	3.7 (2.0)	2.9 (1.7)	3.94	(971)***	0.43
I steal	2.9 (1.7)	2.4 (1.3)	2.46	(968)*	0.33
I tease other kids	3.5 (1.9)	2.8 (1.6)	3.76	(971)***	0.40
I tell lies	3.2 (1.8)	2.7 (1.5)	2.61	(972)**	0.30
I hit kids a lot	3.0 (1.7)	2.5 (1.3)	2.97	(969)**	0.33
It's fun to tease other kids	2.9 (1.7)	2.5 (1.3)	2.74	(972)**	0.26
I like to mess up other kids' games or work	3.0 (1.7)	2.6 (1.4)	2.19	(967)*	0.26
I'm nasty to animals	2.9 (1.6)	2.4 (1.2)	3.13	(971)**	0.35
I fight with other kids a lot	3.1 (1.8)	2.6 (1.4)	3.49	(971)***	0.31
I hit my mummy or daddy	3.1 (1.7)	2.6 (1.3)	2.84	(970)**	0.33
I yell at mummy or daddy	3.0 (1.7)	2.6 (1.4)	2.27	(971)*	0.26
I break other people's things	2.7 (1.5)	2.4 (1.2)	2.38	(969)*	0.22
I start fires	2.8 (1.6)	2.4 (1.2)	3.12	(969)**	0.28
I don't do what mummy/daddy ask me to do	3.1 (1.8)	2.9 (1.6)	1.38	(966)	0.12
I don't do what my teacher asks me to do	3.0 (1.7)	2.5 (1.3)	2.74	(967)**	0.33
I swear or say bad words	3.0 (1.7)	2.5 (1.3)	3.07	(969)**	0.33
I cheat when playing a game	3.1 (1.7)	2.8 (1.6)	1.57	(970)	0.18
BPI total scale	59.0 (15.6)	51.0 (13.0)	5.54	(971)***	0.56

[a] Degrees-of-freedom are based on number of families rather than number of children to account for the dependence in the data due to analyzing two children in the same family (Rogers, 1993; Williams, 2000).

* $p < .05$.** $p < .01$. *** $p < .001$.

associated with IQ measured at age 5, and also with measures of reading and school achievement collected when children were age 7 years. Social disadvantage assessed when children were age 5 years and interviewer's rating of chaos in the home during the age-7 home visit increased along with self-reported conduct problems at age 7 years. Children's self-reports of disruptive behavior were not associated with mothers' or teachers' ratings of internalizing problems, but they were associated with hyperactivity and prosocial behavior according to both mothers and teachers. Compared to children who did not report any conduct problems, children who reported conduct problems at age 7 years had parents who were more antisocial. Scores on all measures of poor parenting quality assessed at age 5 years increased with increasing numbers of disruptive behaviors reported by the children themselves at age 7 years. Interviewers' ratings of negative parenting and parents' neglect during the home visit at age 7 years increased with scores on the Dominic-R. Finally, mothers of children with conduct problems experienced more domestic violence than mothers of children who did not endorse any Dominic-R items. Associations between children's self-reports of conduct problems using the Dominic-R and their correlates were in keeping with correlations reported in previous studies of children's behavioral problems using adults' ratings.

The clinical relevance of the self-report measure of conduct problems at age 7 years was examined by comparing rates of Dominic-R items among children with and without a research diagnosis of conduct disorder by age 7 years (Table IV). Children meeting *DSM-IV* criteria for a research diagnosis of conduct disorder endorsed all Dominic-R items in higher proportion than children without conduct problems. Children who answered positively to any Dominic-R items were between two to six times more likely to have a research diagnosis of conduct disorder by age 7 years compared to those who answered negatively. Moreover, children who said yes to two items or more were four times more likely to have conduct disorder compared to those who endorsed none.

Children's self-reports at age 7 using the Dominic-R were uniquely associated with concurrent adults' ratings of children's conduct problems. Scores on the Dominic-R were linked to teachers' ratings (standardized $\beta = .09$, $t = 3.21$, $p < .001$), over and above mothers' reports ($\beta = .33$, $t = 8.84$, $p < .001$). Similarly, scores on the

Table III. Validity Analysis of the Dominic-R Assessing Conduct Problems at Age 7 Years

	Age-7 children's self-report						
	No symptom		One symptom		Two symptoms or more		
Correlates	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*F*
Independent informants about conduct problems							
Age-5 self-report	50.10 (12.52)	1276	56.38 (15.11)	256	58.22 (16.79)	122	27.35**
Age-7 mothers' report	11.96 (9.69)	1514	14.96 (10.64)	300	19.20 (12.55)	147	20.78**
Age-7 teachers' report	4.42 (7.36)	1405	6.59 (9.25)	276	8.71 (11.36)	140	12.99**
Gender							
Males (%, N) (χ^2)	45.5	701	61.8	177	70.3	93	53.69**
Cognitive abilities							
Age-5 IQ	98.84 (13.94)	1507	95.70 (14.37)	297	95.10 (15.00)	146	7.11**
Age-7 reading score	106.77 (12.77)	1511	105.04 (12.57)	299	102.42 (13.08)	145	6.62**
Age-7 school achievement	3.11 (.90)	1395	2.98 (.98)	276	2.60 (.83)	137	18.83**
Family environment							
Age-5 social disadvantage	1.08 (1.64)	1514	1.39 (1.76)	300	1.52 (1.98)	147	6.32*
Age-7 chaotic home	.97 (1.34)	1507	1.38 (1.50)	300	1.61 (1.70)	146	14.06**
Age-7 internalizing problems							
Mothers' report	7.07 (6.00)	1514	7.50 (6.30)	300	7.50 (5.65)	147	0.74
Teachers' report	5.71 (5.95)	1405	5.79 (6.27)	277	5.44 (5.89)	141	0.15
Age-7 hyperactivity problems							
Mothers' report	8.52 (6.76)	1514	10.33 (7.66)	300	12.55 (7.97)	147	18.05**
Teachers' report	3.84 (5.70)	1403	5.16 (6.63)	276	7.03 (8.09)	141	11.28**
Age-7 prosocial behavior							
Mothers' report	16.57 (3.25)	1514	16.13 (3.23)	300	15.25 (3.67)	147	7.41**
Teachers' report	13.09 (4.72)	1395	11.93 (4.70)	274	11.15 (4.79)	140	10.73**
Parents' antisocial behavior							
Mothers' antisocial behavior	10.60 (9.21)	1513	11.77 (9.31)	298	14.57 (11.34)	146	7.49**
Fathers' antisocial behavior	13.79 (15.74)	1506	17.51 (17.54)	297	19.64 (19.76)	146	8.59**
Parenting quality							
Age-5 corporal discipline	−.08 (1.78)	1498	.29 (1.73)	296	.72 (1.70)	147	13.57**
Age-5 EE negativism	1.42 (.90)	1358	1.57 (.93)	264	1.74 (.98)	134	8.14**
Age-5 EE warmth	3.41 (.94)	1360	3.28 (.97)	264	3.04 (1.09)	134	6.29**
Age-7 negative parenting	.64 (1.48)	1514	1.03 (1.78)	300	1.78 (3.05)	146	11.40**
Age-7 neglect	.62 (1.34)	1509	.82 (1.50)	299	1.51 (2.50)	146	6.37**
Mothers' experience of domestic violence							
Age-5 domestic violence	2.51 (5.48)	1514	3.25 (6.18)	300	3.88 (6.51)	147	4.58*

*$p < .01$. **$p < .001$.

Dominic-R were associated with mothers' reports ($\beta = .15, t = 4.89, p < .001$), beyond information reported by the teachers ($\beta = .32, t = 9.49, p < .001$).

DISCUSSION

We used the Berkeley Puppet Interview and the Dominic-R to test whether children as young as age 5 and 7 years can report valid and useful information about their own conduct problems. Findings showed that children's self-reports of disruptive behavior can be valid when they are gathered with structured and developmentally-appropriate instruments: the two measures were found to be associated with a set of related constructs composed of individual characteristics, children's behavior, and family background. We also showed that children's self-reports of disruptive behavior were useful and provided valuable information: compared to children who did not meet *DSM-IV* criteria, children with a research diagnosis of conduct disorder endorsed in a higher proportion nearly all items captured by the two self-reported instruments, and the self-reported measures also contributed unique information that was not already provided by adults' ratings of children behavioral problems.

Our findings suggest that information provided by children generates similar findings as those obtained using mothers' and teachers' reports. This could be taken to indicate that children's self-reports are a valid measure, but an unnecessary one. Indeed, it can be time-consuming, tedious, and costly to get young children to report about themselves (although the Dominic-R has proven to be an

Table IV. Item Rates of the Dominic-R for Children with and Without a Research Diagnosis of Conduct Disorder (CD) by the Age of 7 Years

Dominic-R items	CD Diagnosis (N =188) %	No CD (N = 1780) %	OR	(95% CI)	d
Do you often cheat?	18.54	8.44	2.47***	(1.51–4.05)	0.50
Have you ever stolen more than once?	8.43	3.06	2.92***	(1.55–5.51)	0.59
Have you ever set a fire on purpose?	3.76	1.45	2.66*	(1.19–5.94)	0.54
Do you often skip school?	4.02	1.47	2.81*	(1.23–6.44)	0.57
Have you ever hurt an animal on purpose?	4.12	1.68	2.51*	(1.06–5.97)	0.51
Have you ever destroyed other people's things on purpose?	8.15	1.42	6.16***	(2.70–14.08)	1.00
Do you often start fights?	20.1	11.13	2.01**	(1.27–3.17)	0.39
Have you ever hurt people on purpose?	10.78	3.04	3.85***	(2.14–6.92)	0.74
Have you ever stolen something right out of somebody's hands?	4.80	1.11	4.50***	(2.04–9.91)	0.83
One symptom on the Dominic-R	15.69	14.53	1.33	(0.85–2.08)	0.16
Two symptoms or more on the Dominic-R	19.47	5.77	4.15***	(2.46–7.00)	0.79

*$p < .05$. **$p < .01$. ***$p < .001$.

inexpensive and quick instrument that is easy to use). Collecting children's own perspectives about their conduct problems may seem like a luxury for researchers with tight grant budgets, and a waste of time for clinicians. However, young children's own perceptions may be an alternative solution to the limits of adults' ratings and observational measures of children's behavior: teachers' nonresponse can considerably limit statistical power or engender additional costs for chasing up; fathers' absence may lead to selectively missing data for children at-risk for behavioral problems; and some observational paradigms still lack construct validity and only aggregate information across a short time-period. Researchers and clinicians gather information from multiple sources of information to compensate the limits of each measure. Our study shows that children can be one of those sources.

The low agreement between children's self-reports and adults' reports, despite the fact that children's self-reports were correlated with related constructs, has three potential explanations. Firstly, adult informants have opportunities to observe different behaviors in specific settings while children have an unrestricted perspective of their own activities. Mothers' and teachers' ratings possibly represent only a partial or limited view of children's behavior. Secondly, children and adults may interpret the same behavior in different ways. Thirdly, adults and children have unique characteristics that influence their reporting skills (e.g., psychopathology, criminal history). Children's and adults' reports can be regarded as documenting different, but valid aspects of children's behavior. All informants' reports are imperfect measures of children's behavior. Our study and others (Hodges, 1993; Loeber, Green, Lahey, & Stouthamer-Loeber, 1989; McConaughy, 2000) suggest that collecting information from multiple sources is necessary for a comprehensive assessment of children's behavior problems.

Some may argue that a strong test of the validity of a measure requires looking at the magnitude of the associations between the measure and its correlates in addition to examining the direction of the associations (Westen & Rosenthal, 2003). We can compare the observed correlations, between children's self-reported disruptive behaviors (using the BPI) and related constructs, with the expected correlation found in the literature employing the same method for measuring the correlates, in the same age group, but with adults' assessment of children's behavior. The correlations between adults' reports of disruptive behavior and gender ($r = .25$; Moffitt, Caspi, Rutter, & Silva, 2001), IQ ($r = -.22$; Lynam, Moffitt, & Stouthamer-Loeber, 1993), family environment ($r = .24$; Bolger, Patterson, & Thompson, 1995), and older children's and adolescents' self-reports ($r = .22$; Achenbach, McConaughy, & Howell, 1987; Essex et al., 2002; van der Ende, 1999) are, in general, low to moderate. These correlations using adults' reports of children's behavior closely matched correlations we reported between children's self-reports of conduct problems and established correlates: .14 for gender, −.25 for IQ, .17 for family environment, and an average of .19 for adults' reports of children's behavior problems. Associations with other correlates of conduct problems reported in the literature may be spuriously inflated because variables, both children's behavior and the correlates having been assessed by the same adult informant; thus we limited our

observations to correlations free of this potential effect, otherwise called shared-method variance.

Our study has some limitations. Firstly, both the BPI and the Dominic-R did not include a fixed reporting period. Interview instruments for young children usually leave out reporting periods because children may have difficulty remembering their activities within a specific time-frame, possibly reducing the reliability of their reports (Schwab-Stone, Fallon, Briggs, & Crowther, 1994). Correlations between informants might be higher if children were asked about behavior that occurred during the same time frame as that asked of mothers and teachers. The lack of a reporting period limits the use of the BPI and the Dominic-R for clinical purposes as they cannot be used alone to establish standardized psychiatric diagnoses. Secondly, it is possible that the validity of children's self-reports varies according to their cognitive abilities. However, further analyses indicated that correlations between self-reported conduct problems and established correlates were similar among children above and below the sample's mean IQ score (available from authors on request). Thirdly, the extent to which our findings generalize to symptoms of other childhood disorders, such as hyperactivity or depression, remains unclear. Indeed, the reliability of children's reports varies depending on the type of disorder children are asked about (Loeber et al., 1989). To limit the length of the interview because of young children's short attention span, E-Risk interviews targeted mainly conduct problems at both ages 5 and 7. Fourthly, this study did not examine the extent of genetic influences on children's self-reports of conduct problems. A previous study from this sample indicated that, similar to adult's reports of children's behavioral problems, scores on the BPI collected at the age-5 assessment were largely influenced by genetic factors (Arseneault et al., 2003). Despite these weaknesses, our study of two developmentally-appropriate instruments is a fair test of the validity and the usefulness of young children's reports about their own conduct problems: we examined a large representative population sample, we used two different age-appropriate instruments at two time points, we collected data from multiple informants, and we scrutinized an extended set of correlates of children's conduct problems.

Our study indicates that young children are valuable informants about their own deviant activities and disruptive behavior: they can report valid information about themselves and they can report useful and unique data not provided by adults. Children can be involved in the investigation of their behavioral problems in the context of research assessments or for clinical purposes.

ACKNOWLEDGMENTS

We are grateful to the Study mothers and fathers, the twins, and the twins' teachers for their participation. Our thanks to Michael Rutter and Robert Plomin for their contributions, to Thomas Achenbach for his permission to adapt the CBCL, and to members of the E-risk team for their dedication, hard work, and insights. We also thank Jennifer Davidson and Barry Milne for their contribution to this manuscript. Dr. Louise Arseneault was supported by the Canadian Institutes of Health Research. Dr. Julia Kim-Cohen was supported by the NIMH Training Program in Emotion Research (T32-MH18931). Professor Terrie E. Moffitt is a recipient of a Royal Society-Wolfson Research Merit Award. The E-Risk Study is funded by the UK Medical Research Council (G9806489).

REFERENCES

Ablow, J. C., & Measelle, J. R. (1999). The Berkeley Puppet Interview (BPI): Interviewing and coding system manuals. Eugene, OR: University of Oregon, Department of Psychology.

Ablow, J. C., Measelle, J. R., Kraemer, H. C., Harrington, R., Luby, J., Smider, N., et al. (1999). The MacArthur Three-City Outcome Study: Evaluating multi-informant measures of young children's symptomatology. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38,* 1580–1590.

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist/ 4-18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1997). *Manual for the Young Adult Self-Report and Young Behavior Checklist*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101,* 213–232.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: Author.

Arseneault, L., Moffitt, T. E., Caspi, A., Taylor, A., Rijsdijk, F. V., Jaffee, S. R., et al. (2003). Strong genetic effects on cross-situational antisocial behaviour among 5-year-old children according to mothers, teachers, examiner-observers, and twins' self-reports. *Journal of Child Psychology and Psychiatry*, *44,* 832–848.

Bennett, N., Jarvis, L., Rowlands, O., Singleton, N., & Haselden, L. (1996). *Living in Britain: Results from the General Household Survey*. London: HMSO.

Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31,* 78–85.

Bolger, K. E., Patterson, C. J., & Thompson, W. W. (1995). Psychological adjustment among children experiencing persistent and intermittent family economic hardship. *Child Development*, *66,* 1107–1129.

Boyle, M. H., Offord, D. R., Racine, Y., Sanford, D., Szatmari, P., Fleming, J. E., et al. (1993). Evaluation of the diagnostic interview for children and adolescents for use in general population samples. *Journal of Abnormal Child Psychology*, *21,* 663–681.

Caldwell, B., & Bradley, R. (1984). *Home Observation for Measurement of the Environment, revised edition*. Homewood, IL: Dorsey.

Caspi, A., Henry, B., McGee, R. O., Moffitt, T. E., & Silva, P. A. (1995). Temperamental origins of child and adolescent behavior problems: From age 3 to age 15. *Child Development*, *66,* 55–68.

Caspi, A., Taylor, A., Smart, M., Jackson, J., Tagami, S., & Moffitt, T. E. (2001). Can women provide reliable information about their children's fathers? Cross-informant agreement about men's antisocial behaviour. *Journal of Child Psychology and Psychiatry*, *42,* 915–920.

Caspi, A., Moffitt, T. E., Morgan, J., Rutter, M., Taylor, A., Arseneault, L., et al. (2004). Maternal expressed emotion predicts children's antisocial behavior problems: Using MZ-twin differences to identify environmental effects on behavioral development. *Developmental Psychology*, *40,* 149–161.

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

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science*, *250,* 1683.

Dodge, K. A., Pettit, G. S., Bates, J. E., & Valente, E. (1995). Social information-processing patterns partially mediate the effect of early physical abuse on later conduct problems. *Journal of Abnormal Psychology*, *104,* 632–643.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development*, *56,* 265–275.

Essex, M., Boyce, T., Goldstein, L., Armstrong, J., Kraemer, H., & Kupfer, D. (2002). The confluence of mental, physical, social, and academic difficulties in middle childhood. II: Developing the MacArthur Health and Behavior Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41,* 588–603.

Flanery, R. C. (1990). Methodological and psychometric considerations in child reports. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 57–82). Boston, MA: Allyn and Bacon.

Goodman, R. (1994). A modified version of the Rutter Parent Questionnaire including extra items on children's strengths: A research note. *Journal of Child Psychology and Psychiatry*, *35,* 1483–1494.

Haddock, C. K., Rinkdskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, *3,* 339–353.

Hodges, K. (1993). Structured interviews for assessing children. *Journal of Child Psychology and Psychiatry*, *34,* 49–68.

Hughes, C., Oksanen, H., Taylor, A., Jackson, J., Murray, L., Caspi, A., et al. (2002). "I'm gonna beat you!" SNAP!: An observational paradigm for assessing young children's disruptive behaviour in competitive play. *Journal of Child Psychology and Psychiatry*, *43,* 507–516.

Keenan, K., & Wakschlag, L. S. (2002). Can a valid diagnosis of disruptive behavior disorder be made in preschool children? *American Journal of Psychiatry*, *159,* 351–358.

Kim-Cohen, J., Arseneault, L., Caspi, A., Polo-Tomas, M., Taylor, A., & Moffitt, T. E. (2005). Validity of DSM-IV conduct disorder in $4\frac{1}{2}$-5-year-old children: A longitudinal epidemiological study. *American Journal of Psychiatry*, *162,* 1108–1117.

La Greca, A. M. (1990). Issues and perspectives on the child assessment process. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 3–17). Boston, MA: Allyn and Bacon.

Lahey, B., Moffitt, T. E., & Caspi, A. (2003). *Causes of conduct disorder and juvenile delinquency*. New York: The Guilford Press.

Landsford, J. E., Dodge, K. A., Pettit, G. S., Bates, J. E., Crozier, J., & Kaplow, J. (2002). Long-term effects of early child physical maltreatment on psychological, behavioral, and academic problems in adolescence: A 12-year prospective study. *Archives of Pediatrics and Adolescent Medicine*, *156,* 824–830.

Loeber, R., Green, S., Lahey, B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviors. *Development and Psychopathology*, *1,* 317–337.

Lynam, D. R., Moffitt, T. E., & Stouthamer-Loeber, M. (1993). Explaining the relation between IQ and delinquency: Class, race, test motivation, school failure or self-control? *Journal of Abnormal Psychology*, *102,* 187–296.

Maynard, R. A. (1997). *Kids having kids: Economic costs and social consequences of teen pregnancy*. Washington, DC: Urban Institute Press.

McConaughy, S. H. (2000). Self-reports: Theory and practice in interviewing children. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 323–352). Bethlehem, PA: School Psychology Program.

Measelle, J. R., Ablow, J. C., Cowan, P. A., & Cowan, C. P. (1998). Assessing young children's views of their academic, social, and emotional lives: An evaluation of the self-perception scales of the Berkeley Puppet Interview. *Child Development*, *69,* 1556–1576.

Moffitt, T. E., Caspi, A., Krueger, R. F., Magdol, L., Silva, P. A., & Sydney, R. (1997). Do partners agree about abuse in their relationships? A psychometric evaluation of interpartner agreement. *Psychological Assessment*, *9,* 47–56.

Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). *Sex differences in antisocial behaviour: Conduct disorder, delinquency, and violence in the Dunedin Longitudinal Study*. Cambridge: Cambridge University Press.

Moffitt, T. E., & the E-Risk Study Team (2002). Teen-aged mothers in contemporary Britain. *Journal of Child Psychology and Psychiatry*, *43,* 727–742.

Murray, L., Woolgar, M., Cooper, P., & Hipwell, A. (2001). Cognitive vulnerability to depression in five-year-old children of depressed mothers. *Journal of Personality and Social Psychology*, *51,* 435–442.

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology*, *20,* 51–63.

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13,* 19–23.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis in behavioral research: Focused comparisons in the analysis of variance*. New York: McGraw-Hill.

Sattler, J. (1992). *Assessment of children: WISC-III and WPPSI-R supplement*. San Diego: Author.

Schwab-Stone, M., Fallon, T., Briggs, M., & Crowther, B. (1994). Reliability of diagnostic reporting for children aged 6–11 years: A test-retest study of the Diagnostic Interview Schedule for Children-Revised. *American Journal of Psychiatry*, *151,* 1048–1054.

Sclare, I. (1997). *The child psychology portfolio*. Windsor, Berkshire: NFER-Nelson Publishing Company.

StataCorp. (2003). Stata statistical software: Release 8.0. College Station, TX: Stata Corporation.

Stone, W., & Lemanek, K. L. (1990). Developmental issues in children's self-reports. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 18–56). Boston, MA: Allyn and Bacon.

Straus, M. A. (1990). Measuring intrafamily conflict and violence: The Conflict Tactics (CT) Scale. In M. A. Straus & R. J. Gelles (Eds.), *Physical violence in American families: Risk factors and adaptations to violence in 8,145 families* (pp. 403–424). New Brunswick, NJ: Transaction.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin TX: PRO-ED.

Trouton, A., Spinath, F. M., & Plomin, R. (2002). Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood. *Twin Research*, *5,* 444–448.

Valla, J., Bergeron, L., Bérubé, H., Gaudet, N., & St-Georges, M. (1994). A structured pictorial questionnaire to assess DSM-III-R-based diagnoses in children (6–11 years): Development, validity, and reliably. *Journal of Abnormal Child Psychology*, *22,* 403–423.

Valla, J., Bergeron, L., Bidaut-Russell, M., St-Georges, M., & Gaudet, N. (1997). Reliability of the Dominic-R: A young child mental health questionnaire combining visual and auditory stimuli. *Journal of Child Psychology and Psychiatry*, *38,* 717–724.

Valla, J., Bergeron, L., & Smolla, N. (2000). The Dominic-R: A pictorial interview for 6–11-year-old children. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39,* 85–93.

van der Ende, J. (1999). Multiple informants: Multiple views. In H. M. Koot, A. A. M. Crijnen, & R. F. Ferdinand (Eds.), *Child psychiatry epidemiology: Accomplishments and future directions* (pp. 39–52). Assen, The Netherlands: Van Gorcum.

Warren, S. L., Oppenheim, D., & Emde, R. N. (1996). Can emotions and themes in children's play predict behavior problems? *Journal of the American Academy of Child and Adolescent Psychiatry*, *35,* 1331–1337.

Webster-Stratton, C. (1998). Preventing conduct problems in Head Start children: Strengthening parenting competencies. *Journal of Consulting and Clinical Psychology*, *66,* 715–730.

Wechsler, D. (1990). *Wechsler Preschool and Primary Scale of Intelligence-Revised*. London: The Psychological Corporation, Harcourt Brace.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84,* 608–618.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56,* 645–646.