Reliability and Validity of Behavior Problem Checklists as Measures of Stable Traits in Low Birth Weight, Premature Preschoolers

Donna Spiker and Helena C. Kraemer
Stanford University

Norman A. Constantine
Far West Laboratory for Educational Research and Development

Donna Bryant
University of North Carolina at Chapel Hill

SPIKER, DONNA; KRAEMER, HELENA C.; CONSTANTINE, NORMAN A.; and BRYANT, DONNA. *Reliability and Validity of Behavior Problem Checklists as Measures of Stable Traits in Low Birth Weight, Premature Preschoolers.* CHILD DEVELOPMENT, 1992, 63, 1481–1496. Mothers, teachers, and assistant teachers completed the Richman Behavior Checklist (BCL) at ages 2 and 3 years and the Achenbach Child Behavior Checklist for Ages 2–3 (CBCL 2–3) at 3 years for a large sample of low birth weight, premature children. Interinstrument correlations for total scores were moderate, higher for teachers and assistant teachers than for mothers, with moderate temporal stability for BCL scores. Interrater agreement for either total scores or classifications of clinically significant scores was moderately high between teachers and assistant teachers only, and children identified as disturbed by mothers versus teachers represent almost nonoverlapping groups. Furthermore, many more children were identified as disturbed using the BCL. The most powerful predictors of mothers' total CBCL 2–3 scores were HOME Inventory scores and self-reported depression. The use of these scales in clinical and research contexts is discussed.

Among clinicians and researchers, there is growing interest in the prevalence, etiology, and developmental course of clinically significant behavioral problems and maladjustment among infants and preschool age children (Campbell, Ewing, Breaux, & Szumowski, 1986; Crowell, Feldman, & Ginsberg, 1988; Earls, 1980; Fagot, 1984; Lerner, Inui, Trupin, & Douglas, 1985; Richman, Stevenson, & Graham, 1982; Wolkind & De Salis, 1982). Early identification and intervention for infants and preschoolers thought to be at risk for emotional and behavioral disorders is especially emphasized in recent legislation (PL 99-457), and there is a growing need for high-quality measures to assist in clinical decision making as well as to assess the effects of early intervention efforts (e.g., Infant Health and Development Program, 1990).

There are now a number of behavior problem rating instruments that are simple to administer to parents and teachers of preschool children (Achenbach, Edelbrock, & Howell, 1987; Behar & Stringfield, 1974; Quay, Hogan, & Peterson, 1987; Richman et al., 1982). Results obtained from these

The Infant Health and Development Program was funded by grants from the Robert Wood Johnson Foundation to the Department of Pediatrics, Stanford University; the Frank Porter Graham Child Development Center, University of North Carolina; and the eight participating universities. Additional support for the National Study Office was provided to the Department of Pediatrics, Stanford University, from the Pew Charitable Trusts; the Bureau of Maternal and Child Health and Resources Development and the National Institute of Child Health and Human Development, HRSA, PHS, DHHS (grant MCJ-060515); and the Stanford Center for the Study of Families, Children, and Youth. We would like to give special thanks to the education directors, teachers, assistant teachers, evaluation coordinators, and infant assessors at the eight sites who trained staff and gathered data for the study. We gratefully acknowledge the helpful comments of three anonymous reviewers, and thank Diane McIntosh for assistance in manuscript preparation. Send correspondence to Donna Spiker, Infant Health and Development Program, Building 460, Stanford University, Stanford, CA 94305.

[*Child Development*, 1992, 63, 1481–1496. 0009-3920/92/6306-0011$01.00]

checklists frequently are treated as if they measure stable characteristics of child behavior. This occurs despite the outstanding questions and limited data available on the reliability and validity of most of these instruments, particularly when used for very young children (Boyle & Jones, 1985; Links, 1983).

Several major relevant issues have been identified. First, most scales include behaviors covering a wide age range, but many behaviors indicative of disorder at one age may be age-appropriate at other ages within that range. Identification of such behaviors is limited by the paucity of good normative data for very young children (Fagot, 1984; Rutter, 1988). Second, there are concerns about poor predictive validity of measures of behavior disorders from early preschool ages to later ages (Campbell et al., 1986; Garrison & Earls, 1985; Stevenson, Richman, & Graham, 1985; Thompson, 1985). Poor prediction suggests either lack of developmental continuity of behavior or the inadequacy of measures to capture that continuity. Given the rapidity of developmental change with very young children, the underlying implicit assumption that behavior problem checklists are capturing stable traits in this age group is questionable and should be subjected to further empirical study.

Third, scales that rely on maternal report may be misleading as the indicator of actual child behavior. Maternally reported behavior problems probably reflect some combination of stable child behavior, context-specific child behavior, and maternal characteristics (Achenbach, McConaughy, & Howell, 1987). This is an especially salient problem with very young children since maternal report is often the sole source of information and/or the clinical referral source.

There is now an abundant literature documenting that the correspondence between maternal reports of behavior problems and assessments of behavioral problems by others has been relatively weak (Achenbach, McConaughy, & Howell, 1987). Correlations between maternal scores on behavior problem checklists and teacher scores and clinicians' diagnoses, for example, are typically low (Gould, Wunsch-Hitzig, & Dohrenwend, 1981; Jensen, Xenakis, Davis, & Degroot, 1988; Kline, 1988). Furthermore, studies with older children have shown that maternal characteristics, such as depression, marital discord, poor social support, and high levels of life stress, are associated with higher maternal reports of behavior problems (Emery, 1982; Goodyer, Wright, & Altham, 1988; Langner, Gersten, & Eisenberg, 1974; Schaughency & Lahey, 1985). These same factors also emerge in recent studies with preschoolers (Jouriles, Pfiffner, & O'Leary, 1988; Webster-Stratton & Hammond, 1988; Wolkind & De Salis, 1982).

Taken together, such findings suggest that interpretations of data from behavior problem checklists must consider the nature of the observer and the context of the child's behavior, as well as the child's behavior per se. Despite increasing evidence that responses on these instruments are associated with characteristics of informants and that responses of different informants who rate children's behaviors in different contexts may produce different results, researchers and clinicians often tend to use these scales as if they are measuring stable characteristics of the children (i.e., characteristics that generalize over context, rater, instrument, and time). This is suggested by the conclusions in research reports in which findings are interpreted as indicative of a certain percentage of disordered children and/or children in need of treatment.

The present study focuses on two widely used preschool behavior problems scales—the Child Behavior Checklist for Ages 2–3 (CBCL 2–3) (Achenbach, Edelbrock, & Howell, 1987; McConaughy & Achenbach, 1988), and the Behavior Checklist (BCL) (Richman, 1977; Richman et al., 1982). These instruments were completed by mothers and early childhood teachers of 2- and 3-year-old low birth weight, premature children who participated in the treatment group of the Infant Health and Development Program (IHDP), a randomized clinical trial of an early intervention program (Infant Health and Development Program, 1990).

Low birth weight, premature infants represent a relevant population for study because of a growing concern about whether this group is at increased risk for behavioral problems. Many recent studies have addressed this question, some with preschoolers (Goldberg, Corter, Lojkasek, & Minde, 1990; Minde et al., 1989), and some with older children (Breslau, Klein, & Allen, 1988; Hawdon, Hey, Kolvin, & Fundudis, 1990; McDonald, Sigman, & Ungerer, 1989; O'Mara & Johnston, 1989; Portnoy, Callias, Wolke, & Gamsu, 1988; Silva, McGee, &

Williams, 1984; Szatmari, Saigal, Rosenbaum, Campbell, & King, 1990).

Some of the issues raised above have been salient in studies with preschoolers. For example, Minde et al. (1989), reporting on a sample of 64 very low birth weight (<1,501 grams), premature 4-year-olds, found that mothers' reports on the BCL resulted in 43% of the sample in the abnormal range, while the corresponding figure for teachers' reports was 24%. In addition, the authors noted that many of the children with scores in the abnormal range had scores that were just over the cut point rather than being extreme. They suggest that this may reflect immaturity in this population rather than true behavioral pathology. Such results highlight the need for caution in interpreting results from behavior problem checklists and for further empirical work to document the reliability and validity of such measures.

Four major research questions are addressed here: (1) *Interinstrument concordance:* Do two different behavior problem instruments yield similar results when used by the same rater at the same age? (2) *Interrater agreement:* Do different raters (mother, preschool teacher, and assistant teacher) using the same instrument rate the children's behavior problems similarly? (3) *Test-retest stability:* Are children rated similarly by the same rater using the same instrument over a 1-year period of time? (4) *Associations with maternal and environmental factors:* What is the association between maternal ratings of behavior problems and teacher ratings together with selected maternal and environmental variables suggested in the research literature? In particular, we have selected a subset of the most frequently cited salient risk factors used by investigators studying behavior problems: maternal depression, marital discord, quality of the caregiving environment, life stress, and maternal education and age (as indicators of caregiving quality and expectations) (Campbell & Ewing, 1990; Egeland, Kalkoske, Gottesman, & Erickson, 1990; Emery, 1982; Goodyer et al., 1988; Gould et al., 1981; Hooks, Mayes, & Volkman, 1988; Lancaster, Prior, & Adler, 1989; Langner et al., 1974; Larson, Pless, & Miettinen, 1988; Pellegrini, 1990; Rutter, 1982; Schaughency & Lahey, 1985; Webster-Stratton & Hammond, 1988). These are factors that either may influence maternal reporting or may be associated in some causal manner with the development of behavior problems and psychopathology generally.

Method

Subjects and Data Collection Procedures

The data reported here are based on the subjects in the Primary Analysis Intervention Group (N = 377) of the IHDP (Infant Health and Development Program, 1990). The IHDP is an eight-site, randomized clinical trial designed to evaluate the efficacy of combining early child development and family support services with high-quality pediatric follow-up in reducing the health and developmental problems among low birth weight, premature infants (birth weight ≤2,500 grams and gestational age ≤37 weeks).

In the IHDP, 985 infants were randomly assigned either to an intervention group or a follow-up group. Further program descriptions are available (Infant Health and Development Program, 1990; Ramey et al., 1992).

The study reported here used data from the intervention group only. The data were collected from the mothers and Child Development Center teachers and assistant teachers when the children were 24 and 36 months (plus or minus 2 weeks) corrected age. Data describing the sample were collected during the neonatal period while the infants were in the nursery immediately after birth and prior to randomization. Table 1 presents a summary of the pertinent data collected throughout the study, the sources, the methods of collection, and when they were collected.

Sample sizes varied somewhat at each data collection point. Teachers and assistant teachers completed behavior problem instruments for those intervention children who were regularly attending the Child Development Center at 24 months and again at 36 months, within a plus or minus 1 month period, designated as active Center participants (i.e., attended at least 3 days per week for the previous month). At 24 months, 290 (76.9%) of the 377 intervention children were active Center participants, and at 36 months, 282 (74.5%) were active Center participants.

Completion of the two different behavior problem instruments at 36 months by the teachers and assistant teachers was done in a counterbalanced random order prescribed by the study investigators from the central office at Stanford University. The instruments were completed on different days within a 1- or 2-week period of time. Teachers and assistant teachers completed the in-

TABLE 1

SUMMARY OF DATA COLLECTED

Instrument/Variable	Source[a]	Method
I. Behavior Problem Scales:		
At 24 and 36 months:		
Richman Behavior Checklist	M	Interview by "blinded" assessor
	T	Self-administered
	AT	Self-administered
At 36 months:		
Achenbach Child Behavior Checklist for ages 2–3	M	Interview by "blinded" assessor
	T	Self-administered
	AT	Self-administered
II. Other maternal/family data:		
At 36 months:		
HOME Inventory Score	M	Interview by "blinded" assessor
General Health Questionnaire Score	M	Interview by nurse or social worker
Life Stress Events Score	M	Interview by nurse or social worker
Marital Quality Score	M	Interview by nurse or social worker
At birth:		
Birth weight	I	Exam in nursery by nurse
Gestational age	I	Ballard exam by nurse or pediatrician
Neonatal Health Index	I	Chart review by nurse
Gender, maternal age, education, ethnicity	M	Interview of mother by nurse

[a] M = mother; T = teacher; AT = assistant teacher; I = infant.

struments independently and did not discuss them with each other until the forms for all children in their classes were completed.

"Blinded" assessors (uninformed about treatment group status) interviewed the mothers using the Richman Behavior Checklist (BCL) during clinic visits at 24 and 36 months, and during a home visit at 36 months using the Achenbach Child Behavior Checklist for Ages 2–3 (CBCL 2–3).[1] Separate structured training programs to insure standard interviewing procedures in gathering the data were conducted with the "blinded" assessors and the teachers and assistant teachers.

Description of Measures

Richman Behavior Checklist (BCL).—This is a 21-item checklist covering 12 areas of behavior yielding one total score (Richman, 1977; Richman et al., 1982). While the available normative data for this instrument are for children over 24 months of age, there are examples of its use with children as young as 24 months (Crowell et al., 1988; McGuire & Richman, 1986).

Achenbach Child Behavior Checklist for Ages 2–3 (CBCL 2–3).—This scale consists of 99 items describing emotional/behavioral problems, and one open-ended item for additional problems (Achenbach, Edelbrock, & Howell, 1987; McConaughy & Achenbach, 1988). The scale yields nine scores: a Total Sum Score, two broad band scores, Internalizing and Externalizing, and six narrow-band scores.

HOME Inventory.—This is a 55-item, semistructured interview measure used to assess the stimulation quality of the home environment for 3–6-year-old children

[1] The interview version of each behavior problem checklist was reviewed and approved by the instrument's developer (Achenbach, Richman) to assure that standardization of administration was maintained. The original response categories were maintained, as was the vast majority of the original wording. Note also that these versions were devised and validated for use by mothers, not teachers.

yielding a total score and eight subscale scores (Caldwell & Bradley, 1984).

General Health Questionnaire.—This is the 12-item version, a self-report scale to measure mental health of the mother. It emphasizes depression and yields one total score (Goldberg, 1972; Goldberg & Huxley, 1980).

Life Stress Events Score.—This scale, developed for this study, is based on the mothers' reports of the presence of 20 stressful life events within the 6 months preceding the 36-month assessment, events based on the two most widely used life stress scales (Holmes & Rahe, 1967; Sarason, Johnson, & Siegel, 1978). A total score from 0 to 20 was obtained.

Marital Quality Score.—This is a 5-item scale modified for the IHDP from a scale by Spanier (1976) for measuring the quality of the marital relationship. It yields one total score from 0 to 19.

Birth weight.—Birth weight was measured to the nearest 10 grams upon admission to the newborn nursery.

Gestational age.—Within the first 48 hours of age, each infant was evaluated using the Ballard assessment (Ballard, Novak, & Driver, 1979). An estimated gestational age was assigned to each infant based on a modification of the Ballard examination (Constantine et al., 1987).

Neonatal Health Index.—This measure of the infant's neonatal health status was developed for the IHDP (Scott, Bauer, Kraemer, & Tyson, 1989). The standard score was set at a mean of 100 and a standard deviation of 16, and was based on the duration of neonatal hospitalization standardized by birth weight, with higher scores indicating better neonatal health.

Infant gender, maternal age, education, and ethnicity.—These data were obtained from maternal interview during the postpartum hospital stay. Ethnicity was recorded as black, Hispanic, other. Maternal education was classified into five levels: less than ninth grade, some high school, high school graduate, some college, college degree or more.

Data Analysis

Concordances were measured and tested using Spearman correlation coefficients for scaled response measures and by Cohen's kappa (Cohen, 1960) for binary response measures. In all cases, random agreement is indicated by a value of zero. Multiple linear regression analysis was used to assess the relation of baseline measures (entered first), teachers' reports of behavior problems (entered second), and familial measures (stepped in last) to maternal reports of behavior problems.

Results

Table 2 shows descriptive statistics for child and maternal demographic baseline information and for the standard instruments used (other than the two behavior problem scales) for children who were active participants in the IHDP Child Development Centers at both 24 and 36 months corrected age ($N = 290$). Varying sample sizes are due to other missing data. This is a representative sample of all inborn low birth weight, premature infants at the eight IHDP study sites (Infant Health and Development Program, 1990).

Table 3 shows the scores on the two different behavior problem scales (BCL, CBCL 2–3) for the three different groups of raters (mothers, teachers, assistant teachers) at the two time points.

For comparative purposes, Richman et al. (1982, p. 100) have reported mean total scores on the BCL for 4-year-olds: 9.7 for a group with clinically rated behavior problems ($N = 94$) versus 6.0 for a control group ($N = 91$).[2] Table 3 shows that all of the mean BCL scores at 24 and 36 months for this low birth weight sample lie between these two means.

Table 3 shows that on the CBCL 2–3 at 36 months, teachers and assistant teachers had comparable mean scores on the total scale and on the two broad band factors, Internalizing and Externalizing, but their scores were lower than the mothers' scores. Achenbach and his colleagues (McConaughy & Achenbach, 1988, p. 37) reported the

[2] Richman (1977) reported mean BCL data for 200 2½–3½-year-olds. For 50 children with a clinical rating of no problems the mean was 6.1 (SD = 2.3). Those with problem ratings of "dubious" to "severe" had mean BCL ratings from 8.1 to 14.0, but no combined mean BCL ratings were reported for all children with clinically rated behavior problems.

TABLE 2

SAMPLE DESCRIPTIVE CHARACTERISTICS

Variable	N^a	%
Gender:		
Male	147	50.7
Female	143	49.3
Maternal education:		
< ninth grade	16	5.5
Some high school	111	38.3
High school graduate	84	29.0
Some college	46	15.9
College degree or more	33	11.4
Maternal ethnicity:		
Black	166	57.2
Hispanic	27	9.3
White, Asian, other	97	33.4
Marital status:		
Married	112	38.6
Single	154	53.1
Divorced/separated	22	7.6
Widowed	2	.7
Father present in home:		
Yes	137	47.2

	N	M	SD	Observed Range	Possible Range
Maternal age (years)	290	24.1	5.8	13–40	. . .
HOME Inventory Score	277	38.4	8.4	11–54	0–55
General Health Questionnaire Score:					
24 months	289	10.1	5.0	0–31	0–36
36 months	285	10.1	4.5	0–27	0–36
Life Stress Event Score	285	4.6	2.2	1–11	0–20
Marital Quality Score	282	9.3	6.4	0–19	0–19
Birth weight (grams)	290	1,825.3	433.5	640–2,500	. . .
Gestational age (weeks)[b]	290	33.2	2.5	25–38	≤37
Neonatal Health Index[c]	290	101.0	15.8	15–136	. . .

[a] Includes subjects who were active participants in the Intervention Child Development Center at 24 months, with non-missing data.

[b] There were three infants with GA = 38 weeks due to recalculation of some gestational ages early in the IHDP study (see Constantine et al., 1987).

[c] This measure of neonatal health status has mean scores set at 100; higher scores represent better health (Scott et al., 1989).

means for total raw scores of mothers of 2- and 3-year-olds as 40.6 for a normative and 70.5 for a clinical sample. Table 3 shows that the total scores for mothers here were comparable to the normative group, while teachers' and assistant teachers' scores were lower. A similar pattern for the three groups of raters was shown for the Externalizing scores. For the Internalizing scores, the teachers' and assistant teachers' scores were similar to the normative group's scores, and the mothers' scores were higher, although not as high as those for the clinical sample.

Table 4 presents data on interinstrument concordance for the two behavior problem scales. All correlations were statistically significant ($p < .0001$) and of a magnitude to suggest that the children's behaviors were rated similarly by the same person using different instruments. Note, however, that the correlations tend to be higher for teachers and assistant teachers than they are for mothers. Using Pearson's test for differences between dependent correlations, for each set of correlations (Total, Externalizing, Internalizing), those between mothers and teachers and between mothers and assistant teachers were significant (all $p < .001$ or .01), but those between teachers and assistant teachers were nonsignificant.

Because the BCL was used at both ages, the stability of ratings over a 12-month pe-

TABLE 3

DESCRIPTIVE STATISTICS OF BEHAVIOR PROBLEM SCORES FROM DIFFERENT INSTRUMENTS AND RATERS

Rater[a]	*N*	*M*	SD	Range
BCL at 24 months: total score:				
M	289	8.8	3.0	2–17
T	290	8.0	3.5	0–20
AT	280	8.1	3.5	1–19
BCL at 36 months: total score:				
M	279	7.7	3.2	0–20
T	281	6.4	3.8	0–17
AT	268	8.8	3.1	3–19
CBCL 2–3 at 36 months: total score:				
M	273	43.1	18.9	3–98
T	281	25.3	21.1	0–98
AT	268	28.6	22.7	0–118
N[b]	273	40.6	19.5	. . .
C[b]	96	70.5	27.2	. . .
CBCL 2–3 at 36 months: internalizing score:				
M	273	9.2	5.6	0–25
T	281	6.8	6.3	0–30
AT	268	7.4	6.6	0–29
N[b]	273	7.9	5.8	. . .
C[b]	96	13.7	9.1	. . .
CBCL 2–3 at 36 months: externalizing score:				
M	273	23.6	10.7	0–49
T	281	14.9	14.4	0–61
AT	268	16.8	15.0	0–66
N[b]	273	22.5	11.8	. . .
C[b]	96	33.2	16.2	. . .

[a] M = mother; T = teacher; AT = assistant teacher; N = normative sample; C = clinical sample.

[b] From McConaughy and Achenbach (1988).

TABLE 4

INTERINSTRUMENT CONCORDANCE AT 36 MONTHS: CORRELATIONS OF BEHAVIOR PROBLEM SCORES FOR DIFFERENT INSTRUMENTS BY THREE DIFFERENT RATERS

Rater[a]	Correlation	*N*
BCL: total score/CBCL 2–3: total score:		
M	.56	272
T	.77	281
AT	.74	268
BCL: total score/CBCL 2–3: externalizing score:		
M	.54	272
T	.76	281
AT	.70	268
BCL: total score/CBCL 2–3: internalizing score:		
M	.48	272
T	.63	281
AT	.62	268

NOTE.—$p < .0001$ for all correlations.

[a] M = mother; T = teacher; AT = assistant teacher.

TABLE 5

INTERRATER AGREEMENT AT 24 AND 36 MONTHS: CORRELATIONS OF BEHAVIOR PROBLEM SCORES FOR DIFFERENT OBSERVERS AND INSTRUMENTS

Rater[a]	Correlation	p Value	N
BCL at 24 months: total score:			
M & T	.08	.19	289
M & AT	.08	.19	279
T & AT	.61	.001	280
BCL at 36 months: total score:			
M & T	.20	.001	279
M & AT	.14	.02	269
T & AT	.66	.001	268
CBCL 2–3 at 36 months: total score:			
M & T	.16	.01	273
M & AT	.07	.27	263
T & AT	.70	.001	268
CBCL 2–3 at 36 months: internalizing score:			
M & T	.16	.01	273
M & AT	.07	.21	263
T & AT	.56	.001	268
CBCL 2–3 at 36 months: externalizing score:			
M & T	.20	.001	273
M & AT	.11	.04	263
T & AT	.76	.001	268

[a] M = mother; T = teacher; AT = assistant teacher.

riod could be assessed. The correlation for the mothers' BCL scores at age 2 and age 3 was .48 ($p < .0001$, $N = 268$). The correlations for the teachers and assistant teachers were calculated only for those children who had the same teacher and assistant teacher, respectively, at the two ages. The correlation for teachers was .57 ($p < .0001$, $N = 151$) and for assistant teachers, it was .52 ($p < .001$, $N = 77$). All correlations indicated moderately high test-retest stability in ratings made at ages 2 and 3.

Table 5 presents data on interrater agreement for different pairs of raters for the total scores of the two instruments and for the two broad-band factor scores for the CBCL 2–3. Overall, the concordance between different observers was moderately large between teachers and assistant teachers (.56–.76) and weak for mothers versus either other group (.07–.20). At 24 months for the total scores on the Richman scale, only the correlation between teachers' and assistant teachers' scores was statistically significant. Note that the magnitude of the correlation between the teachers' and assistant teachers' Externalizing scores was substantially higher than that between their Internalizing scores ($r = .76$ vs. .56; using Pearson's test for dependent correlations, $p < .001$), indicating greater agreement in describing undercontrolling versus overcontrolling behavior.

Table 6 presents data on interrater agreement for classification of clinically significant behavior problems between pairs of observers for the BCL at 24 and 36 months and for the CBCL 2–3 at 36 months. The classifications used cut points provided by the scale developers (i.e., total scores of 10 or above for the BCL, and 64 or above for the CBCL 2–3). Kappa coefficients indicated moderately high agreement between teachers and assistant teachers for both scales for both ages, but not between mothers and either teachers or assistant teachers.

On the CBCL 2–3 at 36 months, mothers rated almost twice as many children in the clinical range as did either teachers or assistant teachers, who had comparable percentages of such ratings (14.7% vs. 7.0% and 7.6%). Of 56 children rated in the clinical range with the CBCL 2–3 by either mothers or teachers, only three (5.4%) were so rated by both groups.

The percentage of children rated in the clinical range was much higher for each of the groups of raters for the BCL than for the CBCL 2–3. Of 118 children rated in the clinical range at 36 months with the BCL by either mothers or teachers, only 21 (17.8%)

TABLE 6

INTERRATER AGREEMENT AT 24 AND 36 MONTHS: PERCENTAGE AGREEMENT ON CLINICAL STATUS OF TOTAL SCORES FOR TWO INSTRUMENTS

INSTRUMENT (*N*)	RATERS' AGREEMENT (%)		
	Percentage Clinical Status[a]		Kappa Coefficient
	Mother	Teacher	
BCL at 24 months (289)	36.7	32.2	.03
BCL at 36 months (279)	26.9	22.9	.01
CBCL 2–3 at 36 months (273)	14.7	7.0	.01
	Mother	Assistant Teacher	
BCL at 24 months (279)	35.8	32.3	.01
BCL at 36 months (269)	27.1	40.2	.05
CBCL 2–3 at 36 months (263)	14.8	7.6	.00
	Teacher	Assistant Teacher	
BCL at 24 months (280)	32.1	32.1	.40
BCL at 36 months (268)	23.5	39.6	.37
CBCL 2–3 at 36 months (268)	7.5	7.8	.34

[a] Clinical cut point: BCL ≥ 10, CBCL 2–3 ≥ 64.

were so rated by both groups. This may reflect differences due to use of different cut points for the two scales or criteria for validating scale cut points or populations used to establish cut points for the two scales. It is also notable that mothers and teachers rated 24 and 16 children, respectively, in the clinical range on both the BCL and the CBCL 2–3, but only two were the same children (2/38 = 5.3%).

Table 7 presents the results of the hierarchical multiple linear regression to predict the mothers' total scores on the CBCL 2–3 at 36 months. Four variables (site, gender, birth weight, and Neonatal Health Index score) were entered first as a block to account for infant neonatal status and site. Next, the average rating of child behavior problems at school (the average total score on the CBCL 2–3 of the teacher and the assistant teacher) was entered. Finally, seven other variables (i.e., maternal education, maternal age, maternal ethnicity, total HOME Inventory score, Marital Quality Score, Life Stress Events Score, maternal depression score) were stepped in if significant.

The background initial status variables accounted for no significant portion of the variance. Further, the teacher's report of behavior problems at school has no significant relation to mothers' reports of behavior problems at home. The most significant predictor of the mothers' scores was the total score on the HOME Inventory, accounting for 13.7% of the variance, indicating that poorer quality of the home environment is associated with higher maternal reports of behavior problems. Maternal depression (as measured by the General Health Questionnaire) also contributed significantly to the prediction: mothers who had higher depression scores reported more behavior problems. The overall model predicted 16.5% of the variance, that is, a multiple correlation coefficient of .41.

Discussion

Since behavior problem rating instruments are currently being widely recommended for use in educational, medical, and research settings to identify preschool-age children who are classified as behaviorally disturbed or at risk (and thus in need of additional evaluation) or who may be in need of early treatment for potential behavioral disorders (Larson et al., 1988; Lerner et al., 1985; McConaughy, 1985; McGuire & Richman, 1986; Rose, Rose, & Feldman, 1989; Thompson, 1985), it is imperative that the quality of such instruments be scrutinized. The data reported in this study corroborate

TABLE 7

STEPWISE LINEAR REGRESSION TO PREDICT MATERNAL REPORT OF CBCL 2–3 TOTAL SCORE AT 36 MONTHS

Variable[a] (N = 271)	R^2	R^2 Increment	p Value
1. Site			. . .
Gender			.5966
Birth weight			.4023
Neonatal health status[b]	.040		.8377
2. Child behavior problems at school[c]	.056	.016	.1328
3. Quality of home environment[d]	.137	.081	.0001
4. Maternal depression[e]	.165	.028	.0038

[a] Variables were entered in the order shown, with no. 1 entered as a block, then no. 2 entered, then other variables entered in the order of their predictive significance. All variables are described in the Method section.

[b] Neonatal Health Index Score.

[c] Average of the teacher's and assistant teacher's total score on the Child Behavior Checklist for Ages 2–3 (CBCL 2–3).

[d] Total score on the HOME Inventory.

[e] Total score on the General Health Questionnaire (GHQ).

and extend concerns raised by other researchers about the interpretation of behavior problem scales in their use with preschool children (Garrison & Earls, 1985; Rutter & Sandberg, 1985; Thompson, 1985). Our results question the assumption that these checklists measure stable traits or characteristics of young children (i.e., characteristics that generalize over context, rater, instrument, and time).

Our results for overall evaluations of children's behavior problems by teachers, using correlations, showed good interinstrument concordance (median .75), good interrater agreement (median .73), and moderate stability over time (median .54). The stable trait so measured may be restricted to child behavior at school, however. The results for maternal reports of children's behavior are less robust in that interinstrument correlation is lower (.56), as is stability over time (.48), than that measured by teachers' reports. Nevertheless, these results are quite respectable as reliability coefficients and are not out of line with results obtained by others.

Problems arise with the interrater concordance results. The lack of correlation between maternal report and teachers' report (.08, .17, .12) cannot be accounted for strictly by the unreliability of either maternal or teacher report. This strongly suggests that these reports are measuring different behavioral characteristics, perhaps reflecting the different contexts in which mothers and teachers are evaluating behaviors as well as the salience of different behaviors in different contexts. The correlation of maternal report with her own characteristics and that of the home environment in the regression analysis, also found by others (Achenbach, McConaughy, & Howell, 1987; Emery, 1982; Gould et al., 1981; Webster-Stratton, 1988), suggests several possibilities as to what those child characteristics might be. It may be that maternal report accurately reflects the child's behavior at home which has little predictive value for behavior at school. It is possible that maternal report reflects instead, or in addition, characteristics of the mother herself or of the home setting.

Home and school environments may elicit and maintain different child behaviors, and different behaviors may be more salient in different settings (e.g., sleep problems at home, peer aggression at school). We conducted secondary analyses on the six narrow-band scales of the 36-month CBCL 2–3 (i.e., Social Withdrawal, Depressed, Sleep Problems, Somatic Problems, Aggressive, Destructive). On every scale, mothers' mean ratings were significantly higher than teachers' mean ratings. This suggests that the mother-teacher discrepancies cannot be largely explained by the salience of different problems in different settings even though there are certain items that will be more salient for mothers in the home context (e.g., resists going to bed at night).

It is possible that the rater herself may elicit different behavior patterns in the child not elicited by others even in the same set-

ting. The same behaviors seen by teachers at school and by mothers at home may be interpreted differently. Age-appropriate behaviors may be reported by mothers as aberrant, but not by well-trained and experienced teachers. Furthermore, some behaviors may be rated as aberrant in one setting and appropriate in another. Regardless of the interpretations made, such discrepancies highlight the possible fluidity of behavior in young children and the measurement difficulties inherent in using rating checklists of behavior problems.

The fact that the BCL identified many more children in the clinical range of behavior problems than did the CBCL 2–3 when used by either mothers or teachers also raises questions about the validity of the scales. It is noteworthy that many of the children identified by the BCL as in the clinical range had scores just over the cut point. That is, of those in the clinical range, those with scores of 10 or 11 represented 37 of 75 children (49.3%) for mothers' reports, and there were 34 of 64 children (53.1%) for teachers' reports. These results are similar to those reported by Minde et al. (1989), who suggested that the BCL may be capturing developmental immaturity rather than behavior disorder in this population.

Achenbach, McConaughy, and Howell (1987), in their meta-analysis of 119 studies of child and adolescence behavior problems, concluded that limited correlations between different informants "poses challenges for clinical assessment" (p. 227), and that "deviance reported by a particular informant does not necessarily mean, however, that the child should be changed. Instead, interventions may sometimes be more appropriately focused on changing the informant's perceptions of or behavior toward the child" (p. 229). Early identification of behaviorally disordered children (i.e., during preschool) may not be very accurate due to the situational nature of young children's behavior or the rater-specific nature of those behaviors as well as the rapidity of developmental change (Thompson, 1985).

The relations between maternal report of behavior problems and the home environment and maternal depression may reflect the lack of validity of such reports for child behavior or may indicate that a poor quality home environment and maternal depression may either be risk factors for behavior problems or represent outcomes of behavior problems. Others have shown similar results (Earls, 1982; Lahey, Russo, Walker, & Piacentini, 1989; Lancaster et al., 1989; Larson et al., 1988; Richman, 1977; Smets & Hartup, 1988; Webster-Stratton, 1988; Wolkind & De Salis, 1982). Even researchers who have reported data suggesting that mothers high in depression may not be unduly distorting their higher reports of behavior problems have cautioned that further empirical data are needed to clarify cross-informant discrepancies (Richters & Pellegrini, 1989).

It is noteworthy that the issues identified here have also plagued research about infant temperament; these are issues about the generalization of such behavior over context, rater, instrument, and time (Bates, 1987). In addition, the recent debate about the conceptual distinctions between temperament and behavior problems in very young children centers around measurement issues similar to those raised here (Bates, 1990; Sanson, Prior, & Kyrios, 1990a, 1990b; Worobey, 1987). These measurement issues also go to the heart of the methodological problems inherent in the empirical study of the origins of psychopathology.

Several limitations of the present study should be noted. First, the use of a 1-year interval for calculating the test-retest stability coefficient for this sample for the BCL included a continuation of an intensive intervention program begun the previous year. One could assume that this might reduce the temporal stability of the ratings if the intervention served to influence the expression of problematic behaviors. From this perspective, the moderate temporal stability on the BCL could be viewed as noteworthy. On the other hand, one could also assume that temporal stability would increase for this sample because all children had a consistent, controlled environment in conjunction with the variety of home environments. In contrast, a general control group would have the same variety of home environments plus the additional variety of extrafamilial environments (e.g., day-care) that are not consistent and controlled, and presumably would be more varied. Although we have no teacher ratings for the IHDP control group, the correlation was .56 ($p < .0001$, $N = 535$) for mothers' ratings on the BCL at 24 and 36 months, quite comparable to the correlation for the treatment group for mothers as well as those for teachers and assistant teachers.

Second, it is not possible to determine whether the rates of clinically significant behavior problems reported by mothers versus

teachers are more accurate, although the finding that maternal reports result in higher rates is consistent with results of other studies mentioned earlier. Again, this may be due to the differing perceptions of the two groups of raters or the differential salience of different behaviors in home versus school environments. In either case, such results cast doubt on the use of behavior problem rating data from a single source as a definitive criterion for screening for behavior disorders in very young children.

Future research could also examine the specific nature of the problematic items endorsed by mothers versus teachers and the number of children in the clinical range with extreme scores to address questions about the kinds and degree of symptoms in this population. It is nevertheless troubling that the two instruments place quite different percentages of children in the clinically problematic range regardless of the group of raters, and that the group of children so identified by mothers and teachers are virtually a nonoverlapping subset of children regardless of the instrument used. The fact that the two instruments used were validated with mothers and not with teachers may account for some of these discrepancies. Yet the higher scores on all six narrow-band scales of the CBCL 2–3 by mothers as compared with teachers suggest that young children will be described differently by these two groups. Those differences may reflect factors related to raters or contexts or some combination of both.

A third limitation of this study is the use of nonstandard measures of marital quality and life stress. Because these variables did not add significant variance to the prediction in the regression model, the results reported do not change. However, it is uncertain whether they would have had predictive value using other measures.

Finally, the study's focus on behavior problem instruments that essentially are inventories of behavior symptoms contrasts with recent conceptualizations about behavioral pathology in very young children that view pathology as relationship disturbances (Sameroff & Emde, 1989). This approach regards the conceptualization of individual psychopathology in very young children as inadequate and possibly inaccurate since the social and emotional behavior of very young children is inextricably bound by the relationships they have with significant adults, particularly caregivers. This approach is quite compatible with the position advanced here, which acknowledges the effects of contexts and raters in summaries of behavioral expression. Furthermore, many of the methodological issues raised here will also affect attempts to define and study "disordered relationships."

Because this study's sample is both biologically and environmentally at risk, these results are especially important. Without a clearer understanding of the meaning of behavior problem ratings in such a population, conclusions about the genuine behavior characteristics of the children remain. Likewise, these results cannot be generalized for ratings made for a population of full-term, socially advantaged preschoolers.

The different policy implications of the different interpretations made from behavior problem ratings data are not trivial. If maternal reports do reflect some combination of behavior problems in the child, risk factors for such behaviors, or outcomes of such behaviors, it would make sense to direct intervention or treatment to the child or to address home environment conditions that affect the child. If, on the other hand, they do *not* reflect problematic behavior in the child, but largely represent problems in the mother's perceptions, intervention or treatment delivered to the child for behavior problems may be unneeded and unlikely to prove effective. Because mothers are more likely to report behavior problems than are school personnel, the identification of at-risk children on the basis of maternal rather than school report might increase the overall investment in prevention and treatment but at the risk of ineffectiveness if the meaning of the maternal report is not fully understood. Thus, the types of early preventive services needed are dependent on an accurate appraisal of the sources of the difficulties.

Ultimately these important issues can only be resolved by obtaining reports on child behavior in the preschool years from multiple sources (teachers, parents, physicians, psychologists), in multiple settings (preschool, home, clinic), with multiple, psychometrically sound instruments and techniques (see Fergusson & Horwood, 1987a, 1987b; Gould et al., 1981; McConaughy, Achenbach, & Gent, 1988; Reich & Earls, 1987). The predictive validity of each of these must be assessed directly by their ability to predict behavioral problems of various kinds later in life (adjustment in entry to school, school behavior, functioning in

school, trouble with the law, etc.). Only then will it be clear which assessments, taken from which respondents, in which settings, and at what ages, serve well to identify those children at risk and in need of prevention or treatment. We hasten to note, however, that the suggestion that is commonly made to use multiple informants and multiple contexts carries with it the additional problem of deciding how to combine or weight the differing information so obtained. We believe that this problem is far from solved, as the data showing that only two children were rated in the clinical range by both mothers and teachers on both the BCL and the CBCL 2–3 indicate.

Recent review articles have urged the routine use of behavior problem checklists in educational (McConaughy, 1985) and medical practice (Eisert, Sturner, & Mabe, 1991; Sturner, 1991; Thompson, 1985). Our results suggest that such clinical applications for toddlers and preschoolers should be treated as experimental until the value of such checklists is clearly established.

Furthermore, we suggest caution in the interpretation of the findings of this study and from the growing body of research using behavior problem rating instruments. These instruments appear to provide information about some combination of the child's behavior, the characteristics of the observer, and the characteristics of the setting. Whether changes in scores on these instruments, or correlations of such scores against other information, reflect changes or correlations of the child's behavior per se or changes or correlations in the respondent or setting remains unclear.

For example, in the randomized clinical trial within which this study was done, the treatment had a statistically significant effect on the maternal report of children's behavior problems using the CBCL 2–3 at 36 months (Infant Health and Development Program, 1990), and very similar results were obtained for the BCL at both 24 and 36 months. A significant treatment by maternal education interaction was found, and the results indicated that the treatment effect was predominantly for less educated mothers. Since the treatment involved not only an educational intervention for the child, but home visits and parent groups designed to educate the mothers, whether treatment changed the behavior of the child, affected how mothers reported behaviors, changed some risk factor in the home or mother for behavior problems in the child, or changed the mother or the home environment is not clear. Because any of these outcomes represents some benefit to the child, in one sense it does not matter. However, not knowing precisely where the benefit lies places a limitation on the understanding of the treatment and the potential for long-term effects important to policy considerations.

There is a complexity inherent in describing and understanding the emotional and social behavior of very young children that has not yet been adequately addressed. What is clear from the present results is that current methods have limitations that must be recognized in order to forestall both poor decision making with respect to individual children in a clinical setting and misleading interpretations in research studies.

References

Achenbach, T. M., Edelbrock, C., & Howell, C. T. (1987). Empirically based assessment of the behavioral/emotional problems of 2- and 3-year-old children. *Journal of Abnormal Child Psychology,* **15**, 629–650.

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

Ballard, J. L., Novak, K. K., & Driver, M. (1979). A simplified method for assessment of fetal maturation of newly born infants. *Journal of Pediatrics,* **95**, 769–774.

Bates, J. E. (1987). Temperament in infancy. In J. D. Osofsky (Ed.), *Handbook of infant development* (2d ed., pp. 1101–1149). New York: Wiley.

Bates, J. E. (1990). Conceptual and empirical linkages between temperament and behavior problems: A commentary on the Sanson, Prior, and Kyrios study. *Merrill-Palmer Quarterly,* 36, 193–199.

Behar, L., & Stringfield, S. (1974). A behavior rating scale for the preschool child. *Developmental Psychology,* **10**, 601–610.

Boyle, M. H., & Jones, S. C. (1985). Selecting measures of emotional and behavioral disorders of childhood for use in general populations. *Journal of Child Psychology and Psychiatry,* **26**, 137–159.

Breslau, N., Klein, N., & Allen, L. (1988). Very low birthweight: Behavioral sequelae at nine years of age. *Journal of the American Academy of Child and Adolescent Psychiatry,* **27**, 605–612.

Caldwell, B. M., & Bradley, R. H. (1984). *Administration manual: Home Observation for*

Measurement of the Environment (rev. ed.). (Available from B. M. Caldwell or R. H. Bradley, College of Education, University of Arkansas, 2801 South University, Little Rock, AR 72204.)

Campbell, S. B., & Ewing, L. J. (1990). Follow-up of hard-to-manage preschoolers: Adjustment at age 9 and predictors of continuing symptoms. *Journal of Child Psychology and Psychiatry*, 31, 871–889.

Campbell, S. B., Ewing, L. J., Breaux, A. M., & Szumowski, E. K. (1986). Parent-referred problem three-year-olds: Follow-up at school entry. *Journal of Child Psychology and Psychiatry*, 27, 473–488.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement*, 20, 37–46.

Constantine, N. C., Kraemer, H. C., Kendall-Tackett, K. K., Bennett, F. C., Tyson, J. E., & Gross, R. T. (1987). Use of physical and neurologic observations in assessment of gestational age in low birth weight infants. *Journal of Pediatrics*, 110, 921–928.

Crowell, J. A., Feldman, S. S., & Ginsberg, N. (1988). Assessment of mother-child interaction in preschoolers with behavior problems. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 303–311.

Earls, F. (1980). Prevalence of behavior problems in 3-year-old children: A cross-national replication. *Archives of General Psychiatry*, 37, 1153–1157.

Earls, F. (1982). Application of DSM-III in an epidemiological study of preschool children. *American Journal of Psychiatry*, 139, 242–243.

Egeland, B., Kalkoske, M., Gottesman, N., & Erickson, M. F. (1990). Preschool behavior problems: Stability and factors accounting for change. *Journal of Child Psychology and Psychiatry*, 31, 891–909.

Eisert, D. C., Sturner, R. A., & Mabe, P. A. (1991). Questionnaires in behavioral pediatrics: Guidelines for selection and use. *Developmental and Behavioral Pediatrics*, 12, 42–50.

Emery, R. E. (1982). Interparental conflict and the children of discord and divorce. *Psychological Bulletin*, 92, 310–330.

Fagot, B. I. (1984). The consequences of problem behavior in toddler children. *Journal of Abnormal Child Psychology*, 12, 385–396.

Fergusson, D. M., & Horwood, L. J. (1987a). The trait and method components of ratings of conduct disorder: Part I. Maternal and teacher evaluations of conduct disorder in young children. *Journal of Child Psychology and Psychiatry*, 28, 249–260.

Fergusson, D. M., & Horwood, L. J. (1987b). The trait and method components of ratings of conduct disorder: Part II. Factors related to the trait component of conduct disorder scores. *Journal of Child Psychology and Psychiatry*, 28, 261–272.

Garrison, W., & Earls, F. (1985). Change and continuity in behavior problems from the preschool period through school entry: An analysis of mothers' reports. In J. E. Stevenson (Ed.), *Recent research in developmental psychopathology* (pp. 51–65). Oxford: Pergamon.

Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. London: Oxford University Press.

Goldberg, D. P., & Huxley, P. (1980). *Mental illness in the community: The pathway to psychiatric care*. London: Tavistock.

Goldberg, S., Corter, C., Lojkasek, M., & Minde, K. (1990). Prediction of behavior problems in 4-year-olds born prematurely. *Development and Psychopathology*, 2, 15–30.

Goodyer, I. M., Wright, C., & Altman, P. M. E. (1988). Maternal adversity and recent stressful life events in anxious and depressed children. *Journal of Child Psychology and Psychiatry*, 29, 651–667.

Gould, M. S., Wunsch-Hitzig, R., & Dohrenwend, B. (1981). Estimating the prevalence of childhood psychopathology: A critical review. *Journal of the American Academy of Child Psychiatry*, 20, 462–476.

Hawdon, J. M., Hey, E., Kolvin, I., & Fundudis, T. (1990). Born too small—is outcome still affected? *Developmental Medicine and Child Neurology*, 32, 943–953.

Holmes, T., & Rahe, R. (1967). The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 11, 213–218.

Hooks, M. Y., Mayes, L. C., & Volkman, F. R. (1988). Psychiatric disorders among preschool children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 623–627.

Infant Health and Development Program. (1990). Enhancing the outcomes of low birth weight, premature infants: A multisite, randomized trial. *Journal of the American Medical Association*, 263, 3035–3042.

Jensen, P. S., Xenakis, S. N., Davis, H., & Degroot, J. (1988). Child psychopathology rating scales and interrater agreement: II. Child and family characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 451–461.

Jouriles, E. N., Pfiffner, L. J., & O'Leary, S. G. (1988). Marital conflict, parenting, and toddler conduct problems. *Journal of Abnormal Child Psychology*, 16, 197–206.

Kline, R. B. (1988). Methodological considerations in the evaluation of the convergence of psychiatric diagnoses and parent-informant

checklists. *Journal of Abnormal Child Psychology,* 16, 289–298.

Lahey, B. B., Russo, M. F., Walker, J. L., & Piacentini, J. C. (1988). Personality characteristics of the mothers of children with disruptive behavior disorders. *Journal of Consulting and Clinical Psychology,* 57, 512–515.

Lancaster, S., Prior, M., & Adler, R. (1989). Child behavior ratings: The influence of maternal characteristics and child temperament. *Journal of Child Psychology and Psychiatry,* 30, 137–149.

Langner, T. S., Gersten, J. C., & Eisenberg, J. G. (1974). Approaches to measurement and definition in the epidemiology of behavior disorders: Ethnic background and child behavior. *International Journal of Health Services,* 4, 483–501.

Larson, C. P., Pless, I. B., & Miettinen, O. (1988). Preschool behavior disorders: Their prevalence in relation to determinants. *Journal of Pediatrics,* 113, 278–285.

Lerner, J. A., Inui, T. S., Trupin, E. W., & Douglas, E. (1985). Preschool behavior can predict future psychiatric disorders. *Journal of the American Academy of Child Psychiatry,* 24, 42–48.

Links, P. S. (1983). Community surveys of the prevalence of childhood psychiatric disorders: A review. *Child Development,* 54, 531–548.

McConaughy, S. H. (1985). Using the Child Behavior Checklist and related instruments in school-based assessment of children. *School Psychology Review,* 14, 479–494.

McConaughy, S. H., & Achenbach, T. M. (1988). *Practical guide for the Child Behavior Checklist and related materials.* Burlington: Department of Psychiatry, University of Vermont.

McConaughy, S. H., Achenbach, T. M., & Gent, C. L. (1988). Multiaxial empirically based assessment: Parent, teacher, observational, cognitive, and personality correlates of child behavior profile types for 6- to 11-year-old boys. *Journal of Abnormal Child Psychology,* 16, 485–509.

McDonald, M. A., Sigman, M., & Ungerer, J. A. (1989). Intelligence and behavior problems in 5-year-olds in relation to representational abilities in the second year of life. *Developmental and Behavioral Pediatrics,* 10, 86–91.

McGuire, J., & Richman, N. (1986). Screening for behaviour problems in nurseries: The reliability and validity of the Preschool Behavior Checklist. *Journal of Child Psychiatry,* 27, 7–32.

Minde, K., Goldberg, S., Perrotta, M., Washington, J., Lojkasek, M., Corter, C., & Parker, K. (1989). Continuities and discontinuities in the development of 64 very small premature infants to 4 years of age. *Journal of Child Psychology and Psychiatry,* 30, 391–404.

O'Mara, L., & Johnston, C. (1989). Mothers' attitudes and their children's behaviors in 3-year-olds born prematurely and at term. *Developmental and Behavioral Pediatrics,* 10, 192–197.

Pellegrini, D. S. (1990). Psychosocial risk and protective factors in childhood. *Developmental and Behavioral Pediatrics,* 11, 201–209.

Portnoy, S., Callias, S., Wolke, D., & Gamsu, H. (1988). Five-year follow-up study of extremely low-birthweight infants. *Developmental Medicine and Child Neurology,* 30, 590–598.

Quay, H. C., Hogan, A. E., & Peterson, D. R. (1987). *Preschool Behavior Problem Checklist (Form II).* (Available from A. E. Hogan, Department of Psychology, P.O. Box 248185, University of Miami, Coral Gables, FL 33124.)

Ramey, C. T., Bryant, D. B., Wasik, B. H., Sparling, J. J., Fendt, K. H., & Lavange, L. M. (1992). The Infant Health and Development Program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics,* 89, 454–465.

Reich, W., & Earls, F. (1987). Rules for making psychiatric diagnoses in children on the basis of multiple sources of information: Preliminary strategies. *Journal of Abnormal Child Psychology,* 15, 601–616.

Richman, N. (1977). Behaviour problems in preschool children: Family and social factors. *British Journal of Psychiatry,* 131, 523–527.

Richman, N., Stevenson, J., & Graham, P. J. (1982). *Pre-school to school: A behavioural study.* London: Academic.

Richters, J., & Pellegrini, D. (1989). Depressed mothers' judgments about their children: An examination of the depression-distortion hypothesis. *Child Development,* 60, 1068–1075.

Rose, S. L., Rose, S. A., & Feldman, J. F. (1989). Stability of behavior problems in very young children. *Development and Psychopathology,* 1, 5–19.

Rutter, M. (1982). Prevention of children's psychosocial disorders: Myth and substance. *Pediatrics,* 70, 883–894.

Rutter, M. (1988). Epidemiological approaches to developmental psychopathology. *Archives of General Psychiatry,* 45, 486–495.

Rutter, M., & Sandberg, S. (1985). Epidemiology of child psychiatric disorder: Methodological issues and some substantive findings. *Child Psychiatry and Human Development,* 15, 209–233.

Sameroff, A. J., & Emde, R. N. (Eds.) (1989). *Rela-*

tionship disturbances in early childhood. New York: Basic.

Sanson, A., Prior, M., & Kyrios, M. (1990a). Contamination of measures in temperament research. *Merrill-Palmer Quarterly,* 36, 179–192.

Sanson, A., Prior, M., & Kyrios, M. (1990b). Further exploration of the link between temperament and behavior problems: A reply to Bates. *Merrill-Palmer Quarterly,* 36, 573–576.

Sarason, I. G., Johnson, J. H., & Siegel, J. M. (1978). Assessing the impact of life changes: Development of the Life Experiences Survey. *Journal of Consulting and Clinical Psychology,* 46, 932–946.

Schaughency, E. A., & Lahey, B. B. (1985). Mothers' and fathers' perceptions of child deviance: Roles of child behavior, parental depression, and marital satisfaction. *Journal of Consulting and Clinical Psychology,* 53, 718–723.

Scott, D. T., Bauer, C. R., Kraemer, H. C., & Tyson, J. (1989). A neonatal health index for preterm infants. *Pediatric Research,* 25, 263a.

Silva, P. A., McGee, R., & Williams, S. (1984). A longitudinal study of the intelligence and behavior of preterm and small for gestational age children. *Developmental and Behavioral Pediatrics,* 5, 1–5.

Smets, A. C., & Hartup, W. W. (1988). Systems and symptoms: Family cohesion/adaptability and childhood behavior problems. *Journal of Abnormal Child Psychology,* 16, 233–246.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marital and similar dyads. *Journal of Marriage and the Family,* 38, 15–28.

Stevenson, J., Richman, N., & Graham, P. (1985). Behavior problems and language abilities at three years and behavioral deviance at eight years. *Journal of Child Psychology and Psychiatry,* 26, 215–230.

Sturner, R. A. (1991). Parent questionnaires: Basic office equipment? *Developmental and Behavioral Pediatrics,* 12, 51–54.

Szatmari, P., Saigal, S., Rosenbaum, P., Campbell, D., & King, S. (1990). Psychiatric disorders at five years among children with birthweights <1,000 g: A regional perspective. *Developmental Medicine & Child Neurology,* 32, 954–962.

Thompson, R. J. (1985). Delineation of children's behavior problems: A basis for assessment and intervention. *Developmental and Behavioral Pediatrics,* 6, 37–50.

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

Webster-Stratton, C., & Hammond, M. (1988). Maternal depression and its relationship to life stress, perceptions of child behavior problems, parenting behaviors, and child conduct problems. *Journal of Abnormal Child Psychology,* 16, 299–315.

Wolkind, S. N., & De Salis, W. (1982). Infant temperament, maternal mental state and child behaviour problems. In R. Porter & G. M. Collins (Eds.), *Temperamental differences in infants and young children* (pp. 221–239). London: Pitman.

Worobey, J. (1987). Temperament ratings in early childhood by parents, teachers, and students. *Early Childhood Research Quarterly,* 2, 169–179.