The MacArthur Three-City Outcome Study: Evaluating Multi-Informant Measures of Young Children's Symptomatology

JENNIFER C. ABLOW, PH.D., JEFFREY R. MEASELLE, PH.D., HELENA C. KRAEMER, PH.D., RICHARD HARRINGTON, M.D., JOAN LUBY, M.D., NANCY SMIDER, PH.D., LISA DIERKER, PH.D., VALERIE CLARK, PH.D., BERNADKA DUBICKA, M.D., AMY HEFFELFINGER, PH.D., MARILYN J. ESSEX, PH.D., AND DAVID J. KUPFER, M.D.

ABSTRACT

Objective: Three sites collaborated to evaluate the reliability and validity of 2 measures, developed in tandem to assess symptomatology and impairment in 4- to 8-year-old children: the Berkeley Puppet Interview Symptomatology Scales (BPI-S) and the Health and Behavior Questionnaire (HBQ). **Method:** In this case-control study, mothers, teachers, and children reported on multiple dimensions of children's mental health for 120 children (67 community and 53 clinic-referred children). **Results:** The BPI-S and the parent and teacher versions of the HBQ demonstrated strong test-retest reliability and discriminant validity on a majority of symptom scales. Medium to strong effect sizes (Cohen *d*) indicated that children in the clinic-referred group were viewed by all 3 informants as experiencing significantly higher levels of symptomatology than nonreferred, community children. **Conclusion:** The availability of a set of multi-informant instruments that are psychometrically sound, developed in tandem, and developmentally appropriate for young children will enhance researchers' ability to investigate and understand symptomatology or the emergence of symptomatology in middle childhood. *J. Am. Acad. Child Adolesc. Psychiatry,* 1999, 38(12):1580–1590. **Key Words:** symptomatology, psychopathology, young children, assessment.

One of the most significant challenges facing the fields of child and adolescent psychiatry and clinical psychology is the accurate measurement of symptomatology and impairment in children younger than 8 years of age. Precise detection of symptomatology and impairment in young children has been impeded by (1) gaps in our understanding of the middle childhood period of development, in particular, forms of impairment that, because they are low in frequency or complex in character, are not detected until the emergence of more pronounced difficulties in mental health (Kazdin, 1994); (2) an absence of age-appropriate methods with which to obtain young children's reports of their own symptomatology; and (3) a lack of integration among measures designed to assess multiple informants' perspectives on young children's mental health.

As part of a larger effort to design a battery of measures to assess biological, neuropsychological, personality, and contextual aspects of development that may contribute to adaptation and impairment during the middle childhood years, the MacArthur Network on Psychopathology and Development has supported the generation of developmentally appropriate assessment methods for children in the 4- to 8-year-old range. Two products of the Network's larger effort, the Berkeley Puppet Interview Symptomatology Scales (BPI-S) and the mental health component of the Health and Behavior Questionnaire (HBQ), were

Accepted July 2, 1999.

Drs. Ablow and Measelle were senior coauthors of this article and, with Dr. Kraemer, are with the Stanford University Medical Center, Stanford, CA. Drs. Harrington, Clark, and Dubicka are with the Royal Manchester Children's Hospital, Manchester, England. Drs. Luby and Heffelfinger are with Washington University School of Medicine, St. Louis. Dr. Dierker is with Wesleyan University, Middletown, CT. Drs. Smider and Essex are with the University of Wisconsin School of Medicine, Madison. Dr. Kupfer is with the University of Pittsburgh School of Medicine.

Oversight for this project provided by Dr. Essex, core member of the John D. and Catherine T. MacArthur Foundation Research Network on Psychopathology and Development. The MacArthur Research Network on Psychopathology and Development, of which Dr. Kupfer is Chair, supported this study.

After January 1, 2000, Dr. Ablow will be with the University of Oregon, Eugene. Requests for reprints and correspondence may be addressed to Dr. Ablow, Department of Psychology, 1227 University of Oregon, Eugene, OR 94703-1227; e-mail: jcablow@oregon.uoregon.edu.

developed in tandem and are the focus of the present investigation. In this article, we present evidence of the BPI-S's and HBQ's internal consistency, test-retest reliability, and ability to discriminate between clinical and nonclinical samples recruited from 3 diverse communities.

An Age-Appropriate Method for Eliciting Children's Reports of Their Own Symptomatology

Eliciting reliable self-reports from young children is complicated by a variety of developmental factors (e.g., short attention spans and less developed language skills) and methodological problems, such as using methods that exceed children's cognitive abilities (e.g., Likert scales) or rely *exclusively* on verbal abilities (for review see Measelle et al., 1998). The Berkeley Puppet Interview (BPI) was developed to address the absence of standardized methods appropriate for young children (Ablow and Measelle, 1993). The BPI is an interactive interview that blends structured and clinical interviewing techniques to elicit children's self-perceptions. Published results demonstrate that the BPI method is a reliable and valid way of obtaining children's perceptions of their competencies and emotional well-being (Measelle et al., 1998). The Berkeley Puppet Interview Symptomatology Scales (BPI-S) were developed explicitly to assess young children's perceptions of their own symptomatology and emotional distress. The interview consists of 9 separate measures of mental health that (1) pertain to symptoms of known importance in this age range, (2) comprise sets of items that can be used to map on to the current diagnostic system for childhood disorders (*DSM-IV*), and (3) were developed to ensure children's comprehension.

Improving the Options for Adult Reports of Symptomatology in Young Children

In parallel fashion, the HBQ was developed to improve the availability of instrumentation for asking parents and teachers about the physical and mental health problems specific to 4- to 8-year-old children. Although measures, such as the Child Behavior Checklist (Achenbach and Edelbrock, 1983) and the Diagnostic Interview Schedule for Children (DISC) (Shaffer et al., 1993), exist to assess the symptomatology of children, the effectiveness of these measures with younger cohorts is not clear; items from these instruments typically are written for older children or to cover a broad age span (e.g., 4–18 years old) (Valla et al., 1994). Given documented links between physical and mental health (Boyce et al., 1995), unlike existing measures, the HBQ also focuses on physical ailments and illnesses that may emerge during this period of childhood. Furthermore, because epidemiological research suggests that child impairment rather than symptom severity tends to be the basis for most clinical referrals (Offord et al., 1996), the HBQ is designed to evaluate young children's adaptation and impairment in addition to their symptomatology. By integrating measures of health, mental health symptomatology, impairment, and social functioning, the HBQ seeks to provide researchers with new ways to organize and understand child well-being in the 4- to 8-year-old range.

Stages of Instrument Development and Aims of This Study

The development and testing of the BPI-S and HBQ were planned in 5 progressive stages: (1) During the theoretical stage the clinical needs of researchers and clinicians working with 4- to 8-year-old children were identified through literature reviews and used to develop drafts of the instruments. (2) During the feasibility stage the instruments were pilot-tested and revised. (3) Stage 3, presented here, was a 3-site case-control study designed to evaluate the psychometric properties of the BPI-S and the mental health component of the HBQ. (4) Now under way, the fourth stage tests the predictive validity of the BPI-S and HBQ in a large community study. (5) Finally, the fifth stage will consist of a large-scale prospective study to evaluate the long-term predictive validity of the BPI-S and HBQ and to develop clinical cutoffs to identify children at risk for subsequent psychopathology.

As stated, this study reports on stage 3 of the 5-stage plan of investigation and was designed to address 3 primary questions. First, are children's responses on the BPI-S and mothers' and teachers' responses on parallel scales from the HBQ internally consistent? Second, do the BPI-S and HBQ demonstrate acceptable test-retest reliability? Third, can these measures discriminate between nonreferred children in the community (low-risk) and children referred to local mental health clinics for evaluation or treatment (high-risk)? The design of this study was based on the premise that if neither instrument could first discriminate between clinical and nonclinical cases, the BPI-S and the HBQ would have even less clinical utility in diagnostically complex samples.

METHOD

Participants

The data reported were obtained during the MacArthur Three-City Outcome Study, a multisite collaboration conducted at 2 sites in the

United States (Palo Alto, California, and St. Louis, Missouri) and 1 site in England (Manchester). These 3 sites were selected for their socioeconomic, ethnic, and cultural distinctiveness. Combined across all 3 sites, 120 children aged 4½ to 7½ years (69 boys and 51 girls), their primary caregivers (95.1% of whom were the biological or adoptive mothers), and their primary teacher or day-care providers participated. Sixty-seven children were recruited from communities near the sites (e.g., pediatric practices, preschool and elementary school settings), and 53 were recruited from local mental health settings (e.g., outpatient clinics and affiliated private practices).

To be included in the study, children had to (1) be between the ages of 4½ years (at least 54 months) and 7½ years of age (at most 90 months), (2) have a legal, primary caretaker (defined here as parent, close relative functioning in loco parentis, legal guardian, or foster parent) with whom they had resided for at least 6 continuous months prior to the study, and (3) have been in a school or day-care setting at least 40 hours prior to the study. In addition, clinic-referred participants had to have been referred for mental health evaluation or treatment prior to enrollment in the study, but had not yet received 1 month of treatment. Boys and girls were excluded if they were currently or recently hospitalized for psychiatric reasons, met diagnostic criteria for severe developmental delays, or had a parent(s) who refused any aspect of the study's procedures. Parents or children who could not speak English with a minimum level of proficiency also were excluded from this investigation.

In total, 158 families expressed interest in the study: 82.7% of the community families and 74.8% of the clinic-referred families agreed to participate; 38 families were not enrolled because of disinterest, time constraints, or eligibility problems.

Measures

Berkeley Puppet Interview Symptomatology Scales. The BPI-S consists of 65 items that were designed specifically to assess children's perceptions of their symptomatology on 9 separate scales: 3 internalizing scales (Overanxious, Separation Anxiety, and Depression), 4 externalizing scales (Oppositional Defiant, Conduct Problems, Overt Hostility, and Relational Aggression), and 2 attention deficit scales (Inattention and Impulsivity).

Interviewers consisted of 1 postbaccalaureate-level, 5 graduate-level, 2 master's-level, and 2 doctoral-level interviewers. All interviewers received BPI administration certification after a standardized, 2-day workshop and the successful completion of 5 to 8 practice interviews (see Ablow and Measelle, 1993). During the actual BPI, children are interviewed with 2 identical, puppy dog hand puppets named Iggy and Ziggy. Throughout the interview, puppets offer opposing statements about themselves and then ask the child, "How about you?" For example, one puppet says, "I'm a sad kid," and the second puppet will say, "I'm not a sad kid. How about you?" All interviews are videotaped for late coding.

Scoring. Based on the degree to which children's responses parallel one of the puppet's statements, responses are coded on a 7-point (1–7) Likert scale, on which very positive self-perceptions or self-reports of no symptomatology (e.g., "I'm *never* a sad kid") are coded on one endpoint of the scale (1) and very negative self-perceptions or reports of severe distress (e.g., "I'm *always* sad") are coded on the other endpoint (7) (see Measelle et al., 1998, for additional information about the BPI's coding system). To test the portability and cross-site reliability of the BPI's coding system, all interviews were scored by 2 coders from different sites. Coders were blind to children's group status. Average interrater agreement across all BPI-S items was 0.87 (Spearman ρ, range = 0.69–1.0).

Health and Behavior Questionnaire. Mothers and teachers completed informant-specific versions of the HBQ, a paper-and-pencil measure that yields dimensional ratings of 4- to 8-year-old children's functioning in 4 domains: (1) emotional and behavioral symptomatology, (2) impairment, (3) adaptive social functioning, and (4) physical health. Although primarily composed of the Ontario Child Health Study Scales (Boyle et al., 1993), the Prosocial Behavior Scale (Weir and Duveen, 1981), and sections of the Medical History Questionnaire from the Rand Health Insurance Study of Children (Lewis et al., 1989), subscales from the Child Behavior Scale (Ladd and Profilet, 1996) and the Preschool Social Behavior Scale (Crick et al., 1997) also were adapted to increase coverage of children's social impairment. Responses on the HBQ are scored on a 3-point Likert scale consisting of 0 ("never or not true"), 1 ("sometimes true"), and 2 ("often or very true"). The HBQ parent version consists of 140 items, and the teacher version consists of 115 items. Both versions take approximately 20 minutes to complete.

This study examined the reliability and validity of the HBQ's symptomatology scales only, which were designed to parallel the BPI's symptomatology scales. Parent and teacher versions of the HBQ are essentially the same, except for minor wording changes (e.g., substituting school for home) and that the teacher version of the HBQ does not assess children's separation anxiety. As such, parents evaluated children's internalizing symptomatology on 3 scales (Overanxious, Depression, and Separation Anxiety), whereas teachers completed only the Overanxious and Depression scales. All adult informants completed 4 externalizing scales (Oppositional Defiant, Conduct Problems, Overt Hostility, and Relational Aggression), and 2 attention deficit scales (Inattention and Impulsivity).

Standardized Assessment of Receptive Language. To determine whether language level influenced how well they understood the BPI, children's receptive language abilities were assessed with the Sentence Structure subtest from the age-appropriate version of the Clinical Evaluation of Language Fundamentals (i.e., depending on their age children completed either the CELF-PRE [Wiig et al., 1992] or the CELF-3 [Semel et al., 1995]). A standard score was computed for each child by using age norms provided by the test manufacturer.

Procedures

After eligibility was established during an initial telephone screener, families were scheduled for 2 home visits separated by no fewer than 7 and no more than 10 days. Before the first home visit, parents were sent a packet that contained the HBQ and a demographics questionnaire. After their first visit, parents were informed if they had been randomly selected for the retest visit.

Research teams consisting of a child interviewer and a parent interviewer visited families in their homes for approximately 1½ to 2 hours. After consent was obtained from the parent and assent was obtained from the child, parent and child worked independently with their respective interviewer. Administration of the puppet interview was broken into 2 parts and separated by a break. During the break, the child received a snack and completed the receptive language assessment. Home visits ended with 15 minutes of free play. While the child participated in the puppet interview, the parent was interviewed with the computerized DISC (DISC-IV). Although not included in the present investigation, the DISC-IV's categorical data will be contrasted in future reports with the HBQ's dimensional approach to evaluating young children.

Although all families were scheduled for a second visit, only half of the sample was randomly selected for retesting. Of the 60 families selected for retesting, 56 successfully completed the second visit; 4 families were unable to keep their retest appointment and were randomly replaced to maintain power. All aspects of the study were readministered during the retest visit, except for the DISC-IV and the CELF

To avoid biasing the results, a different research team completed the retest interviews. For participating in the study, parents received $25 and children received a small gift after each home visit.

Once signed parental consent was received, teachers were mailed an assessment packet that included (1) an instructional cover letter and study description, (2) a copy of the corresponding family's release form, (3) the HBQ-Teacher Version, and (4) an addressed, stamped envelope to return the completed HBQ. Teachers were paid $25 for each HBQ they completed.

RESULTS

Description of Sample

Demographic characteristics of the entire sample, broken down by site and group status, were analyzed first. At all 3 sites, the community and clinical groups did not differ with respect to child gender or mean age (5.9 years, SD = 0.98). Sites did differ significantly in terms of ethnicity, $\chi^2(12, N = 120) = 131.23, p < .0001$, with the Palo Alto (13.6%) and St. Louis (17.4%) sites having some minority representation (specifically, African-American, Asian, or Latino families) and Manchester having only white participants. The community and clinical samples differed significantly on receptive language ability, with both preschool-age children (mean = 10.8, SD = 3.0), $F_{2,56} = 11.86, p < .01$, and school-age children (mean = 12.3, SD = 1.9) from the community, $F_{2,66} = 4.52, p < .04$, scoring higher on the CELF than clinic-referred children (means = 7.9 and 10.1, SDs = 3.1 and 2.1, respectively). Despite differences in receptive language ability, children in both groups scored in the normal range and had sufficient receptive language skills to complete the BPI.

Sites differed significantly in terms of household income, $F_{2,109} = 16.36, p < .0001$, with the Palo Alto sample (mean = $94.5, SD = $14.2) reporting greater household income than the St. Louis sample (mean = $64.6, SD = $23.2) and both U.S. sites reporting greater household income than the Manchester sample (mean = $50.5, SD = $12.7). Groups also differed significantly on household income, with families in the community sample reporting greater income (mean = $77.7, SD = $16.3) than families in the clinical sample (mean = $62, SD = $20.3), $F_{1,109} = 4.23, p < .05$.

Internal Consistency of the BPI and HBQ Scales

Coefficient α values are presented in Table 1 separately by informant and group status. In all but 5 instances, the internal consistency of children's responses on the BPI-S exceeded 0.60, and most α coefficients exceeded .70. Although modest by adult standards, these α values are consistent with other studies using child informants (Hodges, 1993; Measelle et al., 1998) and can be considered acceptable given the breadth of symptoms covered by each scale. As shown in Table 1, the internal consistency of the scales completed by mothers and teachers was high. Finally, although the differences were modest, the α coefficients of most scales completed by or about children in the clinic-referred sample exceeded the α coefficients in the community sample.

Test-Retest Reliability

Table 2 provides information on each instrument's test-retest reliability (and 95% confidence intervals). Reliability coefficients (product-moment *r* values) were based on the correlation between scale means that were collected 7 to 10 days apart. Children's assessments of their own symptom-

TABLE 1
Internal Consistency of Rationally Constructed Scales for Each Informant Measure and by Group

	BPI-Child		HBQ-Parent		HBQ-Teacher	
Scales by Sample	No. of Items	α	No. of Items	α	No. of Items	α
Depression						
Community	7	.36	7	.79	6	.75
Clinical	7	.75	7	.68	6	.56
Overanxious						
Community	7	.62	12	.77	8	.71
Clinical	7	.77	12	.78	8	.77
Separation Anxiety[a]						
Community	6	.63	10	.57	—	—
Clinical	6	.71	10	.86	—	—
Oppositional Defiant						
Community	6	.62	9	.76	9	.84
Clinical	6	.71	9	.85	9	.91
Conduct Problems						
Community	9	.51	12	.79	11	.79
Clinical	9	.84	12	.88	11	.87
Overt Hostility						
Community	7	.73	4	.63	4	.78
Clinical	7	.89	4	.77	4	.88
Relational Aggression						
Community	6	.71	6	.84	6	.92
Clinical	6	.68	6	.86	6	.92
Inattention						
Community	5	.58	6	.75	6	.90
Clinical	5	.69	6	.83	6	.83
Impulsivity						
Community	6	.52	9	.82	9	.87
Clinical	6	.71	9	.86	9	.90

Note: BPI = Berkeley Puppet Interview; HBQ = Health and Behavior Questionnaire.

[a] Separation anxiety items were not included in the teacher version of the HBQ.

TABLE 2
Test-Retest Reliability (Pearson *r*) and Confidence Intervals for Each Informant Measure and by Group

Scales	BPI-Child[a]		HBQ-Parent[a]		HBQ-Teacher[b]	
	r	95% CI	*r*	95% CI	*r*	95% CI
Depression						
Community	0.43	(0.14–0.67)	0.82	(0.67–0.91)	0.88	(0.77–0.94)
Clinical	0.42	(0.05–0.70)	0.83	(0.65–0.92)	0.88	(0.74–0.95)
Separation Anxiety[c]						
Community	0.58	(0.32–0.76)	0.80	(0.64–0.89)	—	—
Clinical	0.60	(0.28–0.80)	0.96	(0.91–0.98)	—	—
Overanxious						
Community	0.72	(0.52–0.85)	0.80	(0.64–0.89)	0.83	(0.68–0.91)
Clinical	0.69	(0.42–0.85)	0.87	(0.73–0.94)	0.83	(0.64–0.93)
Oppositional Defiant						
Community	0.59	(0.33–0.77)	0.83	(0.69–0.91)	0.86	(0.73–0.93)
Clinical	0.61	(0.29–0.82)	0.78	(0.56–0.90)	0.96	(0.91–0.98)
Conduct Problems						
Community	0.60	(0.35–0.78)	0.87	(0.76–0.93)	0.83	(0.68–0.91)
Clinical	0.69	(0.42–0.85)	0.83	(0.65–0.92)	0.90	(0.78–0.96)
Overt Hostility						
Community	0.52	(0.24–0.73)	0.80	(0.64–0.89)	0.74	(0.53–0.86)
Clinical	0.54	(0.20–0.78)	0.89	(0.77–0.94)	0.93	(0.84–0.97)
Relational Aggression						
Community	0.53	(0.24–0.73)	0.77	(0.59–0.88)	0.88	(0.77–0.94)
Clinical	0.70	(0.44–0.86)	0.87	(0.73–0.94)	0.88	(0.74–0.95)
Inattention						
Community	0.56	(0.29–0.75)	0.81	(0.66–0.90)	0.87	(0.75–0.93)
Clinical	0.65	(0.36–0.83)	0.61	(0.29–0.81)	0.95	(0.88–0.98)
Impulsivity						
Community	0.49	(0.20–0.71)	0.79	(0.62–0.89)	0.88	(0.77–0.94)
Clinical	0.76	(0.54–0.89)	0.81	(0.59–0.91)	0.93	(0.64–0.93)
Mean of all scales[d]						
Community	0.59	(0.33–0.77)	0.80	(0.64–0.89)	0.81	(0.64–0.89)
Clinical	0.62	(0.28–0.83)	0.81	(0.59–0.91)	0.90	(0.78–0.96)

Note: BPI = Berkeley Puppet Interview; HBQ = Health and Behavior Questionnaire; CI = confidence interval. *R* values (2-tailed) are statistically significant if 0 does not fall within the confidence interval.

[a] For children and mothers, community $n = 36$ and clinic-referred $n = 27$.

[b] For teachers, community $n = 35$ and clinic-referred $n = 25$.

[c] Separation anxiety items were not included in the teacher version of the HBQ.

[d] Fisher *r* to *z* transformation.

atology demonstrated moderately high reliability, averaging 0.60 for the entire sample (95% confidence interval = 0.47–0.71). Comparisons between the clinical and community values indicated that the reliability coefficients of children in the clinical sample tended to exceed those of children in the community sample (66% of the time, $p < .05$ by one-tailed sign test); however, the magnitude of these differences was not statistically significant. Except for the Depression scale, community and clinical children were equally reliable when reporting on different aspects of their symptomatology.

Overall, test-retest reliability was high for mothers and teachers, averaging 0.80 and 0.85, respectively (95% CI = 0.72–0.88 and 0.75–0.90, respectively). Differences between mothers' and teachers' reports were not statistically significant and they demonstrated similar reliability across all domains.

Discriminant Validity Analyses

To test each measure's discriminant validity, we used a 3-way analysis of variance with group (community and clinical), site (Manchester, Palo Alto, and St. Louis), children's gender, and their interactions as factors. The results of these analyses are presented in Table 3, along with the adjusted means, pooled standard deviations, and effect sizes.

Six of the 9 BPI symptomatology scales differentiated between children on the basis of their group status (Table 3). Specifically, the Overanxious, Depression, Oppositional

TABLE 3

Discriminant Validity Analysis of Individual Scales: Group by Site by Child Gender Analyses of Variance of Each Informant's Mean Score

	Manchester				Stanford				St. Louis						
	Boys		Girls		Boys		Girls		Boys		Girls				Group
Scale	COM	CLN	COM	CLN	COM	CLN	COM	CLN	COM	CLN	COM	CLN	(SD)[a]	Significant Main & Interaction Effects	Effect Size[b]
Child															
Depression	2.92	3.80	2.69	3.32	2.91	3.01	2.62	3.14	2.84	2.96	2.71	3.50	(0.92)	G**	0.55
Overanxious	3.81	3.74	2.90	2.82	3.00	3.62	3.00	3.70	2.84	3.46	2.92	4.20	(1.00)	G* S×N**	0.53
Separation Anxiety	3.91	4.00	3.84	3.67	3.34	3.86	3.17	3.58	3.23	3.84	2.85	0.45	(1.16)	G*	0.45
Oppositional Defiant	2.85	3.38	2.24	2.82	2.70	3.24	2.61	2.53	2.66	3.48	2.74	3.34	(0.85)	G**	0.57
Conduct Problems	2.67	2.48	2.24	2.00	2.50	3.03	2.57	2.24	2.36	2.51	2.34	2.90	(0.69)	G* N** G×N* S×N**	0.68
Overt Hostility	2.71	4.13	2.00	2.00	2.35	3.00	2.51	2.19	2.16	2.46	2.44	2.36	(0.86)	G* N** G×N** S×N**	0.40
Relational Aggression	3.25	4.02	2.27	2.00	2.57	3.37	2.03	2.94	2.71	2.85	2.51	2.59	(0.80)	N*** G×N** S×N**	0.36
Inattention	3.93	3.85	2.86	3.20	3.37	3.92	3.49	3.22	2.92	3.93	3.11	4.62	(1.18)	G×S**	0.38
Impulsivity	4.10	3.73	3.22	2.62	3.49	3.10	3.55	3.11	3.06	4.00	3.02	4.76	(1.00)	G×S** S×N*	0.20
Mother															
Depression	0.14	0.58	0.17	0.46	0.20	0.43	0.26	0.71	0.33	0.33	0.13	0.41	(0.28)	G***	0.93
Overanxious	0.33	0.76	0.58	0.83	0.35	0.48	0.43	0.65	0.37	0.63	0.33	0.56	(0.30)	G***	0.87
Separation Anxiety	0.23	0.72	0.46	0.68	0.33	0.54	0.40	0.52	0.31	0.48	0.31	0.38	(0.35)	G**	0.60
Oppositional Defiant	0.55	1.36	0.47	1.21	0.44	0.89	0.53	1.08	0.78	1.01	0.47	0.84	(0.36)	G*** G×S*	1.44
Conduct Problems	0.20	0.89	0.03	0.84	0.15	0.45	0.13	0.55	0.28	0.61	0.13	0.31	(0.30)	G*** G×S**	1.50
Overt Hostility	0.44	0.25	0.21	0.31	0.41	0.81	0.31	1.00	0.47	0.86	0.23	0.55	(0.41)	G*** G×S**	1.51
Relational Aggression	0.52	0.98	0.17	0.71	0.15	0.42	0.43	0.53	0.37	0.41	0.30	0.36	(0.39)	G** S×N*	0.62
Inattention	0.72	1.46	0.70	1.33	0.45	1.24	0.56	0.94	0.46	1.23	0.37	0.85	(0.41)	G***	1.54
Impulsivity	0.94	1.69	0.79	0.25	0.45	1.12	0.67	1.09	0.69	1.43	0.56	0.90	(0.39)	G*** S** G×N* S×N*	1.44
Teacher															
Depression	0.06	0.58	0.13	0.22	0.25	0.65	0.07	0.67	0.23	0.46	0.14	0.36	(0.28)	G***	1.25
Overanxious	0.32	0.50	0.58	0.17	0.42	0.56	0.22	0.79	0.33	0.67	0.36	0.41	(0.35)	G**	0.50
Separation Anxiety[c]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Oppositional Defiant	0.49	0.90	0.22	0.67	0.16	0.82	0.12	0.69	0.25	0.61	0.21	0.54	(0.42)	G***	1.21
Conduct Problems	0.26	0.66	0.02	0.50	0.03	0.50	0.04	0.27	0.09	0.27	0.10	0.15	(0.28)	G*** N*	1.11
Overt Hostility	0.36	1.03	0.08	0.75	0.25	0.96	0.06	0.50	0.31	0.64	0.21	0.38	(0.45)	G*** N**	1.09
Relational Aggression	0.44	0.70	0.42	0.79	0.10	0.61	0.36	0.67	0.08	0.06	0.39	0.40	(0.44)	G** S**	0.52
Inattention	0.57	1.20	0.47	0.92	0.25	1.25	0.22	0.50	0.31	1.26	0.39	0.56	(0.46)	G*** N** G×N*	1.24
Impulsivity	0.70	1.20	0.19	1.00	0.39	1.12	0.20	0.61	0.42	1.25	0.39	0.58	(0.46)	G*** N***	1.24

Note: COM = community sample; CLN = clinic-referred sample; G = group; S = site; N = child gender; G×S = group-site interaction; G×N = group-gender interaction; S×N = site-gender interaction. Means are not broken down by gender, but tabled means have been adjusted by group, site, and gender.

[a] Estimate of pooled within-group standard deviation.

[b] Cohen *d*.

[c] Separation anxiety items were not included in the teacher version of the Health and Behavior Questionnaire.

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

Defiant, and Conduct Problems scales produced moderate group effect sizes, whereas the Separation Anxiety and Overt Hostility scales produced small group effect sizes, according to Cohen's (1988) distinction between small (0.20 through 0.49), moderate (0.40 through 0.79), and large (≥0.80) effect sizes. Although the means for both groups tended to be low, children in the clinic group consistently reported higher levels of symptomatology than their community counterparts. In addition, there were significant gender effects for the Conduct Problems and Overt Hostility scales, with boys reporting higher levels of symptomatology than girls. Two scales, the Inattention and Impulsivity scales, did not produce significant group effects but did yield significant group-site interactions. Single-*df* contrasts revealed that these interactions were due primarily to the clinical sample at St. Louis, where significantly higher levels of Inattention, $F_{1,119} = 9.49, p < .01$, and Impulsivity, $F_{1,119} = 18.32, p < .0001$, were reported than for all other groups and sites combined. Indeed, the group effect sizes for these 2 scales were large at St. Louis, 0.72 and 1.02, respectively.

Mothers' and teachers' HBQ scores are also presented in Table 3. A significant group effect was found with all HBQ scales. Both mothers and teachers consistently reported that children in the clinical group had significantly higher levels of symptomatology than children in the community sample. All group effect sizes for mothers and teachers were moderate to large, with the effect sizes in the externalizing domain consistently exceeding 1.0.

Intercorrelation and Latent Structure of Informant Scales

Next we explored the dimensionality and underlying structure of the BPI-S and HBQ scales. In light of the large number of correlations for each source, comments are limited to the clearest and most encompassing patterns for each measure. When the Fisher *z* transformation of the Pearson *r* values was used to compute an estimate of the average intercorrelation among the scales, the data indicated that children's, mothers', and teachers' responses were moderately intercorrelated (overall *r* values = 0.33, 0.44, and 0.45, respectively). Scales in the externalizing or disruptive behavior domain were the most highly intercorrelated for children, mothers, and teachers (average *r* values = 0.45, 0.66, and 0.68, respectively). In the internalizing domain, children's, mothers', and teachers' reports were only modestly intercorrelated (overall *r* values = 0.28, 0.33, and 0.45, respectively). Across symptom domains, the overall association between the externalizing and internalizing scales was weak to modest for children, mothers, and teachers (*r* values = 0.28, 0.34, and 0.24, respectively).

Next we conducted separate principal component analyses (PCAs) of each informant's scale scores to evaluate further the measure's latent structure. Although an oblique rotation yielded equivalent solutions, we report on the results of a varimax rotated PCA given that our goal was to develop measures that maximally discriminate between domains of symptomatology. The initial results produced identical 3-factor solutions for children's self-reports and teachers' ratings. The rotated solutions for children and teachers produced an externalizing factor, an internalizing factor, and a third factor composed of the Inattention and Impulsivity scales. The initial PCA with the mothers' data yielded a 2-factor solution, composed of an externalizing factor that included the Inattention and Impulsivity scales and an internalizing factor. However, because the Inattention (0.58) and Impulsivity (0.53) scales loaded moderately on the externalizing factor, we forced a 3-factor solution on the maternal data. Not surprisingly, the 3-factor solution yielded a third factor that comprised the Inattention and Impulsivity scales.

Final Test of Discrimination and Cross-Informant Agreement at the Syndrome Level

On the basis of the results of the PCAs, we created 3 composite scales that paralleled the 3-factor solutions just described. Unit weights were used to create syndrome scale scores. These syndrome scales were labeled *Externalizing* (consisting of the Oppositional Defiant, Conduct Problems, Overt Hostility, and Relational Aggression scales), *Internalizing* (consisting of the Overanxious, Separation Anxiety, and Depression scales), and *Attention Deficit* (consisting of the Inattention and Impulsivity scales). The intercorrelations among the 3 syndrome scales were modest, ranging from 0.39 to 0.42 for the children, 0.48 to 0.63 for the mothers, and 0.36 to 0.60 for the teachers.

As shown in Table 4, the syndrome scales provided stronger group discrimination than the individual scales, especially for children's reports. In contrast to their component scales, the Internalizing and Externalizing scales produced large effect sizes. The children's Attention Deficit scale yielded a small group effect size and a significant group-site interaction indicating that the clinical sample at St. Louis reported significantly higher levels of attentional problems than all other children. As expected, each of the syndrome scales for mothers and teachers yielded large effect sizes.

TABLE 4
Discriminant Validity Analysis of Syndrome Scales: Group by Site by Child Gender Analyses of Variance of Each Informant's Mean Score

	Manchester				Stanford				St. Louis						
	Boys		Girls		Boys		Girls		Boys		Girls				
Scale	COM	CLN	COM	CLN	COM	CLN	COM	CLN	COM	CLN	COM	CLN	(SD)[a]	Significant Main & Interaction Effects	Group Effect Size[b]
Child															
Internalizing	3.54	4.01	3.14	3.92	3.08	3.64	2.93	3.52	2.97	3.42	2.83	4.05	(0.79)	G**	0.84
Externalizing	2.71	3.73	2.16	3.19	2.51	3.08	2.56	2.62	2.39	2.79	2.51	2.81	(0.70)	G** N** G×N** S×N*	0.89
Attention Deficit	3.98	3.69	2.99	2.91	3.43	3.51	3.72	3.17	3.00	3.97	3.06	4.88	(0.97)	G×S*** S×N*	0.30
Mother															
Internalizing	0.23	0.69	0.40	0.66	0.29	0.48	0.37	0.65	0.36	0.48	0.25	0.45	(0.24)	G***	1.00
Externalizing	0.41	1.12	0.22	0.98	0.28	0.63	0.35	0.77	0.48	0.73	0.29	0.50	(0.30)	G*** G×S**	1.50
Attention Deficit	0.83	1.58	0.74	1.29	0.45	1.18	0.61	1.02	0.58	1.33	0.46	0.87	(0.37)	G*** S*** N* G×N*	1.62
Teacher															
Internalizing	0.21	0.56	0.39	0.24	0.33	0.69	0.15	0.70	0.30	0.70	0.25	0.47	(0.28)	G***	1.04
Externalizing	0.39	0.82	0.18	0.68	0.14	0.72	0.14	0.53	0.19	0.39	0.23	0.37	(0.36)	G***	1.03
Attention Deficit	0.64	1.20	0.33	0.96	0.32	1.19	0.21	0.56	0.37	1.25	0.39	0.57	(0.42)	G*** S*** G×N*	1.36

Note: COM = community sample; CLN = clinic-referred sample; G = group; S = site; N = child gender; G×S = group-site interaction; G×N = group-gender interaction; S×N = site-gender interaction. Means are not broken down by gender, but tabled means have been adjusted by group, site, and gender.

[a] Estimate of pooled within-group standard deviation.

[b] Cohen *d*.

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

A comprehensive reporting on the patterns of cross-informant agreement is beyond the goal and scope of this article. We provide, however, a brief summary of these data. Young children's reports of their externalizing symptomatology were significantly related to mothers' ($r = 0.44$, $p < .001$) and teachers' ratings ($r = 0.47$, $p < .001$), and the level of agreement between mothers and teachers was considerable ($r = 0.67$, $p < .0001$). In the internalizing domain, children's reports on the BPI-S were not related to teachers' reports ($r = 0.16$, not significant) but were significantly, albeit modestly, associated with maternal report ($r = 0.27$, $p < .01$). The level of concordance between mothers' and teachers' ratings of children's internalizing symptoms was significant, although modest ($r = 0.34$, $p < .001$). Finally, children's reports of their attentional problems were unrelated to their mothers' reports ($r = 0.05$, not significant) but were somewhat related to teachers' reports ($r = 0.23$, $p < .05$). Mothers' and teachers' ratings of children's attentional problems were highly related ($r = 0.72$, $p < .0001$).

DISCUSSION

Internal Consistency

The purpose of this investigation was to examine the internal consistency, test-retest reliability, and discriminant validity of the HBQ and the BPI-S in a case-control study of 4½- to 7½-year-old children. Half of the children were recruited from the community and the other half had been referred for mental health evaluations. The results from this study suggest that when responding to the HBQ, parents and teachers produce internally consistent reports of children's symptomatology. Likewise, except for community-recruited children's responses to items that assess depressive symptoms, the internal consistency of young children's reports on the BPI-S were acceptable and stronger than previous attempts to ask children within this age range to evaluate their mental health (see Byrne, 1996).

The low internal consistency of community children's responses to items from the Depression scale may be due to a lack of experience with or recognition of depressive symptoms. Research on childhood mood disorders suggests that young children who have experienced depression tend to be more sophisticated about and better able to identify their feelings of sadness, loneliness, and self-worth (Richters, 1992). Furthermore, in comparison with the other scales, the BPI's Depression scale is less behaviorally grounded. Whereas the other scales ask directly about

actual experiences (e.g., "My teacher tells me to sit down a lot"), the Depression scale relies on children's subjective sense of their own affect that are independent of context (e.g., general feelings of sadness and loneliness). Improved internal consistency might be achieved with items that assess children's perceptions of experientially based emotions (e.g., sadness that results from peer rejection).

For all 3 informants, the α coefficients in the clinical sample tended to be stronger than the α coefficients in the community sample. Although not intended as diagnostic instruments, the HBQ and BPI-S scales were designed to assess the range of symptomatology needed to make diagnostic decisions. Consequently, some of the items have low base rates, whereas other items could be part of most children's daily experiences. Despite face validity, items with low base rates (e.g., "I start fires when a grown-up is not around") can lower internal consistency, especially in community samples. Likewise, items endorsed by many children can also contribute to decreased internal consistency. For example, the item "Sometimes I cheat" may be normative for children in this age group. Thus, a community-recruited child may endorse that he cheats sometimes, yet look inconsistent when he disavows other items on the Conduct Problems scale. Given that the goals of the HBQ and BPI-S scales are symptom breadth rather than homogeneity, the internal consistency results suggest that all 3 informants perceive more consistent evidence of symptomatology among the clinic-referred children than among the community sample.

Test-Retest Reliability

Mothers' and teachers' reports on the HBQ demonstrated strong test-retest reliability for both the clinic-referred and community children. Furthermore, in contrast to prior studies in which adults' assessments of externalizing disorders are more reliable in older than younger children, and more reliable than adult assessments of internalizing disorders (Fallon and Schwab-Stone, 1994), in the present study, the reliability of adults' reports were comparably high across all HBQ domains.

Children's reports of their own symptomatology on the BPI-S demonstrated impressive test-retest reliability across most domains and were consistently better than the reliability estimates that are often obtained with adult-styled interviews adapted for young children (see Hodges, 1993). Although not significantly different, children in the clinical group provided more stable reports of their symptomatology than children in the community sample. When combined with the higher levels of internal consistency, these results may reflect that the symptoms of clinic-referred children are more stable than those of their community counterparts. It would follow then that clinic-referred children would be more reliable and aware of these symptoms that either they or others have identified and labeled for them. Alternatively, lower reliability in the community sample can be attributed to low symptom base rates (Fallon and Schwab-Stone, 1994).

Discriminant Validity of the HBQ and the BPI-S

Results of this study demonstrated that all of the parent and teacher HBQ scales provided significant and meaningful discrimination between the nonreferred and clinic-referred samples of young children (i.e., larger effect sizes). Although previous studies designed to compare groups of older children and adolescents have been able to differentiate between inpatients, outpatients, and community comparisons (Costello et al., 1985), to date few empirical examples of teachers' and parents' reports on children in the 4- to 8-year-old range have demonstrated the level of discrimination produced by the HBQ in this study.

In contrast to prior difficulties eliciting meaningful information directly from young children, in this study, children's scores on a majority of the BPI-S scales exhibited significant discriminant validity. This study demonstrates that young children can provide valid self-reports when interviewed with an age-appropriate method. Furthermore, whereas earlier studies have shown that older children and adolescents tend to be better reporters of their internalizing than externalizing symptomatology (Hinshaw et al., 1992), in this study, young children's reports discriminated in both domains. Specifically, the clinic-referred children at all 3 sites reported higher levels of depression, overanxiousness, and separation anxiety than the community sample. In the externalizing domain, clinic-referred children reported greater levels of oppositionality, conduct problems, and overt aggression than their community counterparts.

The Inattention and Impulsivity scales of the BPI-S failed to yield a significant main effect for groups, yet they did produce a significant site by group interaction. Post hoc chart reviews revealed that 82% of the clinical sample at St. Louis had been referred specifically for attention-deficit/hyperactivity disorder evaluations, in contrast to 49% of the clinical samples at Manchester and 34% at Palo Alto. Analysis of this interaction revealed that the clinical sample at St. Louis presented with significantly

higher levels of impulsivity and inattention than the other clinic-referred and community participants. While these scales will require additional development given that they failed to yield significant group effects at 2 of the 3 sites, we are encouraged by the validity of the Attention Deficit scales in a sample in which there was ample evidence of problems that may meet criteria for attention-deficit/hyperactivity disorder or attention deficit disorder.

Discrimination at a Broad-Band Level of Psychopathology

For each measure, separate factor analyses of the 9 symptomatology scales yielded similar 3-factor solutions that consisted of a primary externalizing factor, a secondary internalizing factor, and a weaker, though coherent factor pertaining to problems with attention. The results of this study closely parallel the empirically derived, broad-band dimensions of child psychopathology reported by Achenbach and his colleagues (Achenbach et al., 1987). Furthermore, results demonstrate that for mothers, teachers, and children, these composite scales provided better discrimination than the individual symptom scales. The improved discrimination can be attributed partially to gains in reliability that are achieved when combining symptom clusters, such as depression, overanxiousness, and separation anxiety, that tend to be part of the same constellation of childhood problems (Cohen, 1988). Children in this sample were included in the clinic-referred group on the basis of their risk status (i.e., referral for mental health assessment or treatment), independent of specific symptom presentations or actual diagnosis. However, had we sampled by diagnostic classification, the discriminant validity of the HBQ's and BPI-S's individual syndrome scales may have been stronger.

Limitations

A central goal of this study was to increase generalizability by using a multisite design. Although this objective was met on several important levels (e.g., findings replicated across site, family socioeconomic status, informant, and child gender), we fell short in terms of the sample's overall ethnic diversity. In planning a large-scale replication or extension of this type of design, specific recruitment procedures would need to be implemented to ensure that different ethnic and cultural groups were adequately represented. In constructing a larger replication sample, attention should also be paid to children's specific diagnostic classification. The results reported herein were purposely limited to young children who had and had not been referred for mental health evaluations or treatment; we considered this a necessary first step in evaluating the BPI-S's and HBQ's utility as clinical research tools. Although neither the BPI-S nor the HBQ is intended to yield discrete diagnoses, an important follow-up would be to examine the correspondence between our dimensional measures and children's diagnoses as indicated by the DISC-IV.

Clinical Implications

The past decade has been a time of remarkable growth and progress in the study of child psychopathology. The assessment of mental health in young children, however, is far less developed than procedures for determining mental health status in older children and adolescents. Due to a dearth of methods validated specifically for children in the middle childhood years, the field continues to lack clarity about the nature of psychopathology in this age period. Without well-founded, age-appropriate means to ask parents, teachers, and children about mental health, many young children presenting with early evidence of emerging psychopathology or "below threshold" symptomatology are not identified as targets for intervention until their symptoms become significantly worse or exacerbated. It is our hope, however, that the availability of a set of multi-informant instruments, such as the BPI-S for children and the HBQ for parents and teachers, that are psychometrically sound, developed in tandem, and specific to 4- to 8-year-old children, will enhance researchers' ability to investigate factors relevant to the emergence and course of early psychopathology.

REFERENCES

Ablow JC, Measelle JR (1993), *Berkeley Puppet Interview: Administration and Scoring System Manuals.* Berkeley: University of California

Achenbach TM, Edelbrock CS (1983), *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington: University of Vermont Department of Psychiatry

Achenbach TM, McConaughy SH, Howell CT (1987), Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull* 101:213–232

Boyce WT, Chesney M, Alkon-Leonard A et al. (1995), Psychobiologic reactivity to stress and childhood respiratory illnesses: results of two prospective studies. *Psychosom Med* 57:411–422

Boyle MH, Offord DR, Racine YA, Fleming JE, Szatmari P, Sanford M (1993), Evaluation of the revised Ontario Child Health Study scales. *J Child Psychol Psychiatry* 34:189–213

Byrne BM (1996), *Measuring Self-Concept Across the Life Span: Issues and Instrumentation.* Washington, DC: American Psychological Association

Cohen J (1988), *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum

Costello EJ, Edelbrock CS, Costello AJ (1985), Validity of the NIMH Diagnostic Interview Schedule for Children: a comparison between psychiatric and pediatric referrals. *J Abnorm Child Psychol* 13:579–595

Crick NR, Casas JF, Mosher M (1997), Relational and overt aggression in preschool. *Dev Psychol* 33:579–588

Fallon T, Schwab-Stone M (1994), Determinants of reliability in psychiatric surveys of children aged 6–12. *J Child Psychol Psychiatry* 35:1391–1408

Hinshaw SP, Han SS, Erhart D, Huber A (1992), Internalizing and externalizing behavior problems in preschool children: correspondence among parent and teacher ratings and behavior observations. *J Clin Child Psychol* 21:143–150

Hodges K (1993), Structured interviews for assessing children. *J Child Psychol Psychiatry* 34:49–67

Kazdin AE (1994), Informant variability in the assessment of childhood depression. In: *Handbook of Depression in Children and Adolescents,* Reynolds WM, Johnston HF, eds. New York: Plenum, pp 249–271

Ladd GW, Profilet SM (1996), The Child Behavior Scale: a teacher-report measure of young children's aggressive, withdrawn, and prosocial behavior. *Dev Psychol* 34:267–283

Lewis M, Pantell RH, Kieckhefer GM (1989), Assessment of children's health status: field test of new approaches. *Med Care* 27:54–64

Measelle JR, Ablow JC, Cowan PA, Cowan CP (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 Dev* 69:1556–1576

Offord DR, Boyle MH, Racine Y et al. (1996), Integrating assessment data from multiple informants. *J Am Acad Child Adolesc Psychiatry* 35:1078–1085

Richters JE (1992), Depressed mothers as informants about their children: a critical review of the evidence for distortion. *Psychol Bull* 112:485–499

Semel E, Wiig EH, Secord WA (1995), *Clinical Evaluation of Language Fundamentals,* 3rd ed. San Antonio, TX: Psychological Corporation

Shaffer D, Schwab-Stone M, Fisher P et al. (1993), The Diagnostic Interview Schedule for Children-Revised Version (DISC-R), I: preparation, field testing, interrater reliability, and acceptability. *J Am Acad Child Adolesc Psychiatry* 32:643–650

Valla J-P, Bergeron L, Berube 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 reliability. *J Abnorm Child Psychol* 22:403–423

Weir K, Duveen G (1981), Further development and validation of the Prosocial Behavior Questionnaire for use by teachers. *J Child Psychol Psychiatry* 22:357–374

Wiig EH, Secord W, Semel E (1992), *Clinical Evaluation of Language Fundamentals-Preschool.* San Antonio, TX: Psychological Corporation

---

Communicating Bad News: A Pediatric Department's Evaluation of a Simulated Intervention. Larrie W. Greenberg, Daniel Ochsenschlager, Regina O'Donnell, Jennifer Mastruserio, George J. Cohen

Objective: To determine if pediatric residents and emergency department (ED) fellows could improve their ability to counsel and inform standardized patients (SPs) about bad news. *Methodology:* A crossover, self-controlled design in which trainees were their own control individuals, and SPs provided feedback after the first interview. The setting was the consultation room in the ED of a large children's hospital. The outcome measures included examining the counseling and informing skills of study participants. *Results:* Trainees improved their informing skills after being provided feedback in the broad areas of communication and follow-up and in the total number of content areas asked. Their counseling skills improved in two areas: 1) promoting more trust and 2) making parents feel less dependent. Those trainees who scored higher on counseling skills discussed more total and critical content issues with SPs in the study. Trainee feedback revealed a very high rating of the educational process, and the trainees also felt much more confident about their skills after the first and second sessions. *Conclusions:* Using SPs to teach residents and ED fellows to give bad news is an effective educational process that provides trainees with interactions that simulate real-life experience. **Pediatrics** 1999;103: 1210–1217. Reproduced by permission of *Pediatrics,* copyright 1999.