Test-Retest Reliability of the Preschool Age Psychiatric Assessment (PAPA)

HELEN LINK EGGER, M.D., ALAATTIN ERKANLI, Ph.D., GORDON KEELER, M.Sc.,
EDWARD POTTS, Ph.D., BARBARA KEITH WALTER, Ph.D., M.P.H.,
AND ADRIAN ANGOLD, MRCPsych.

ABSTRACT

Objective: To examine the test-retest reliability of a new interviewer-based psychiatric diagnostic measure (the Preschool Age Psychiatric Assessment) for use with parents of preschoolers 2 to 5 years old. **Method:** A total of 1,073 parents of children attending a large pediatric clinic completed the Child Behavior Checklist 1½–5. For 18 months, 193 parents of high scorers and 114 parents of low scorers were interviewed on two occasions on an average of 11 days apart. **Results:** Estimates of diagnostic reliability were very similar to those obtained from interviews with parents of older children and adults, with κ ranging from 0.36 to 0.79. Test-retest intraclass correlations for *DSM-IV* syndrome scale scores ranged from 0.56 to 0.89. There were no significant differences in reliability by age, sex, or race (African American versus non-African American). **Conclusions:** The Preschool Age Psychiatric Assessment provides a reasonably reliable standardized measure of *DSM-IV* psychiatric symptoms and disorders in preschoolers for use in both research and clinical service evaluations of preschoolers as young as 2 years old. *J. Am. Acad. Child Adolesc. Psychiatry,* 2006;45(5):538–549. **Key Words:** preschoolers, psychopathology, reliability, community sample, diagnostic interview.

Empirical studies of emotional and behavioral difficulties in infants and preschoolers have been few compared with studies of older children and adolescents. Until recently, studies of psychopathology in very young children have tended to concentrate on parent/caregiver-reported symptom and scale scores derived from

symptom checklists, rather than the standardized diagnostic categories of the *DSM-IV-TR* (American Psychiatric Association, 2000) or the International Classification of Diseases-10 (World Health Organization, 1993; for a review see Angold and Egger, 2004). The information derived from checklist measures such as the Child Behavior Checklist 1½–5 (CBCL/1½–5; Achenbach and Rescorla, 2000) does not provide details of severity, frequency, or duration on the range of symptoms needed to enable researchers or clinicians to make the sorts of psychiatric diagnoses that we are familiar with at every other stage of life. The psychometric properties of instruments such as the CBCL/1½–5 (Achenbach and Rescorla, 2000) are encouraging and at the very least indicate that relatively stable psychopathological characteristics can be identified in preschoolers.

A surprisingly small group of studies has assessed the applicability of DSM psychiatric diagnoses to young children. These studies have used unmodified interviews developed for use with older children including the Diagnostic Interview Schedule for Children (DISC; Briggs-Gowan et al., 2000; Speltz et al., 1995, 1999) or

the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS; Keenan et al., 1997; Lahey et al., 1998; Shaw et al., 1996, 2001); sections of these interviews modified to be developmentally appropriate for younger children such as a modified DISC depression module (Luby et al., 2002a, 2003) or modified disruptive behavior disorders modules of the K-SADS (Keenan and Wakschlag, 2000, 2002); or unstructured clinical interviews (Kashani et al., 1984; Kashani et al., 1986; Kashani et al., 1997) or developed diagnoses based on clinical consensus (e.g., Earls, 1982; Lavigne et al., 1996).

A few studies have used diagnoses defined in the Diagnostic Classification: 0–3 (DC: 0–3; Zero to Three, 1994), an alternative psychiatric diagnostic classification for young children. They relied on unstructured clinical interviews (Guedeney et al., 2003; Keren et al., 2003; Reams, 1999; Thomas and Clark, 1998) or semistructured clinical assessments (Boris et al., 1998; Cordeiro et al., 2003; Scheeringa et al., 1995, 2003).

Although no published psychometric data are available on the use of any of these measures with preschool children, it does appear that instruments or approaches designed and tested for use with older children can be applied to younger children and that they yield total prevalence estimates similar to those found for older children and adolescents (Earls, 1982; Keenan et al., 1997; Lavigne et al., 1996). However, there is also great variation in the apparent rates of specific diagnoses from study to study, disagreement over which diagnostic system to use, and little evidence that any instrument is really performing appropriately.

The exception to this lack of attention to the psychometrics of preschool diagnosis occurs in the assessment of autism and pervasive developmental disorders (PDDs; Volkmar et al., 2004). The use of the Autism Diagnostic Interview (Lord et al., 1994), a parent-report structured interview, and the Autism Diagnostic Observation Schedule (Lord et al., 1989) in numerous studies during the last decade has led to the standardization of the assessment of PDDs, refinement of the phenotype of autism and PDDs, and advances in our understanding of the pathophysiology and genetics of PDDs. The track record of the Autism Diagnostic Interview demonstrates that it is possible to develop a reliable and valid diagnostic assessment of preschool psychiatric symptoms and that such instruments are critical for advancing the understanding of the prevalence and etiology of the disorders.

Exploration of the presentation, course, and outcomes of the full range of preschool psychopathology will be inhibited until we have demonstrably reliable and well-validated structured diagnostic measures that assess the full range of psychiatric symptoms and disorders in young children. Here we present data from the first test-retest reliability study of a parent-report diagnostic instrument for the comprehensive assessment of preschool psychopathology, the Preschool Age Psychiatric Assessment (PAPA; Egger and Angold, 2004; Egger et al., 1999).

METHOD

Description of the PAPA

Development of the PAPA began in 1998 and has been described in some detail elsewhere (Egger and Angold, 2004). A brief summary is presented here. The PAPA is a parent-report instrument based on the parent version of the Child and Adolescent Psychiatric Assessment (CAPA) for 9- to 18-year-olds (Angold and Costello, 1995, 2000b; Angold et al., 1995). Table 1 details the key features of the PAPA, including training, modes of administration, and the content of the PAPA modules. Trained PAPA interviewers interview the parent or other primary caregiver. The interviewer codes the PAPA after the interview is completed. If the paper version of the PAPA is used, then the individual PAPA variables are then entered into a database. If the ePAPA, the electronic version of the PAPA, is used, then the interview and coding are completed on a tablet PC, making separate data entry unnecessary. In this study, we used a paper version of the PAPA, PAPA 1.3. Symptoms, diagnoses, and scale scores were generated from the raw interview symptom data by a series of computerized algorithms developed in SAS (SAS, Cary, NC).

Like the CAPA, the PAPA combines the characteristics of an interviewer-based and a respondent-based interview. Like respondent-based interviews, the PAPA uses a highly structured protocol, with required questions and probes. However, as in an interviewer-based interview, the onus throughout is on the interviewer to ensure that subjects understand the question being asked, provide clear examples about behaviors or feelings relevant to the symptom, and have the symptom at a prespecified level of severity as defined in an extensive glossary. Whenever a symptom is endorsed, the interviewer must write down examples. If the interviewer determines that a symptom is present, the frequency, duration, and dates of onset of the symptoms are separately assessed. Data on the relationship and setting context of the symptom are also collected. A 3-month "primary period" is used rather than a longer period because shorter recall periods are associated with more accurate recall (Angold et al., 1996a). Lifetime occurrence of severe symptoms including suicidality, fire-setting, and cruelty to animals and people are assessed, as well as lifetime occurrence of potentially traumatic life events such as child abuse or death of a parent.

Although the overall conceptualization, design, and implementation of the PAPA will be familiar to users of the CAPA, it nonetheless involves significant revisions of content and structure to

TABLE 1
Structure and Content of the PAPA

Interview	Comprehensive, structured, glossary-based psychiatric interview for assessing psychiatric symptoms, symptom scale scores, and diagnoses, as well as life events, family structure and functioning, and impairment in preschool children
Informant	Parent or other primary caregiver
Age range	2–5 yr (24–72 mo)
Interviewers	Trained lay interviewers (having at least a bachelor's degree) or mental health professionals
Training	5 days for entire interview plus 2–3 practice interviews; required for all interviewers
Translations	Spanish (under development), Romanian, Norwegian
Methods of administration	Paper version of the PAPA and electronic version (ePAPA) administered on a tablet PC
Content	25 modules; individual modules can be administered separately or in any combination:
Modules	Brief Developmental Assessment
	Family Structure and Functioning; Parental Psychopathology
	Childcare
	Play and Peer Relationships
	Depression
	Mania
	Conduct Problems
	Attention-Deficit/Hyperactivity Disorder
	Tics and Tricotillomania
	Stereotypies and Unusual Speech (screening for PDDs)
	Regulation/Habits
	Eating and Food-Related Behaviors
	Sleep
	Somatization
	Elimination
	Separation Anxiety
	Anxious Affect
	Worries
	Rituals and Repetitions
	Psychosis
	Reactive Attachment Disorder Symptoms
	Life Events
	Posttraumatic Stress Syndrome
	Incapacity/Disability/Impairment
	Socioeconomic Status and other demographic information

make it relevant for the assessment of younger children. The PAPA assesses symptoms in four domains. The first is assessment of all *DSM-IV-TR* and ICD-10 diagnostic criteria insofar as they are relevant to this age group. Examples of developmentally inappropriate items that were excluded from the PAPA include questions about substance use, sexual history, and some conduct problems including committing truancy, stealing cars, and breaking curfew. The second item is all items in the Research Diagnostic Criteria-Preschool Age (RDC-PA; Task Force on Research Diagnostic Criteria: Infancy and Preschool, 2003). Sponsored by the American Academy of Child and Adolescent Psychiatry, the RDC-PA Task Force, a group of researchers in preschool and infant psychiatry, proposed these developmentally appropriate revisions of *DSM-IV-TR* diagnostic criteria based on current data and clinical consensus. The third domain is items in the DC: 0–3, both the original version published in 1994 and the revised version (DC: 0–3R) published in August 2005 (Zero to Three, 2005). The fourth domain is potentially

relevant behaviors and symptoms experienced by preschoolers and their families that are not explicitly included in current diagnostic criteria. For example, we developed comprehensive sections on sleep behaviors (e.g., bedtime rituals, place of sleep initiation, behaviors interfering with sleep initiation, nap history), feeding history and eating behaviors, toileting history and elimination patterns, play and peer relationships, and daycare and school settings and experiences. Questions to screen for PDDs are included, but the PAPA does not attempt to diagnose autism spectrum disorders or PDDs because reliable and valid structured measures for these disorders already exist.

A key point about the PAPA is that most sections include some behaviors that are regarded as being normal in preschoolers at certain levels of frequency and pathological at other levels of frequency (e.g., temper tantrums, impulsivity). Because population-based norms for many preschool behaviors and symptoms are lacking, it is important to be able to assess the prevalence, frequency, duration, content, and context of such behaviors to provide

epidemiologic data on the prevalence and distribution of these behaviors and then empirically determine the boundaries between normative and pathological or "clinically significant" behaviors. There are also substantial developmental changes across the preschool period (e.g., Rothbart et al., 2003), and the PAPA provides the ability to define age-specific diagnostic criteria.

Scoring and Diagnostic Algorithms

Diagnostic algorithms were written for the following diagnoses: attention-deficit/hyperactivity disorders (ADHD; inattentive type, hyperactive-impulsive type, combined type), oppositional defiant disorder (ODD), conduct disorder (CD), depression (major depression, dysthymia and depression-not otherwise specified), anxiety disorders (separation anxiety disorder [SAD], generalized anxiety disorder [GAD], specific phobia, social phobia, posttraumatic stress disorder [PTSD], and selective mutism), and elimination disorders (enuresis and encopresis). Composite diagnoses were also developed. Behavioral disorders included having one or more diagnoses of ADHD, ODD, and CD. Emotional disorders included the presence of one or more diagnosis of depression or an anxiety disorder. Any diagnosis included behavioral and/or emotional disorders but does not include the elimination disorders. Scale scores consisting of counts of the number of diagnostic criteria met for each of the diagnoses were also created.

Like the CAPA, the PAPA separately assesses the presence of the symptoms and the presence of disabilities (impairment in *DSM* terms) resulting from symptoms. Here we use the World Health Organization's International Classification of Functioning, Disability and Health (ICF; World Health Organization, 2001) definition of disabilities as negative functional outcomes resulting from health conditions, involving significant deviation from or loss of normal or expected function (Angold and Costello, 2000a). The PAPA assesses disability in 30 areas including the child's relationships with his or her parents, other adults, siblings, and peers, as well as the child's functioning in the home (e.g., inability to carry out simple chores, inability to dress oneself), at school, or in daycare (e.g., being suspended or expelled from school or daycare), and out of the home (e.g., inability to go to a grocery store with a parent, eat at a restaurant, or attend a religious service). By separately assessing the effect of symptoms on functioning and on the quality of the child's relationships with significant others, one can distinguish between functional impairment and distress caused by the symptoms. Disability was considered present if the parent reported that the child was disabled in one or more areas. A disability scale consisting of the number of areas of impairment (scale ranges from 0–30) was generated. We also created a variable for "serious emotional disturbance" (SED), the term used by the U.S. government (*Federal Register*, 1993) for psychiatric disorder accompanied by significant impairment in the child's functioning.

Table 2 outlines the diagnostic criteria used for the PAPA algorithms. Diagnoses included the *DSM-IV-TR* (and RDC-PA) frequency and duration criteria, as well as onset criteria when indicated. As far as possible, the PAPA algorithms followed the CAPA algorithms developed for older children. However, modifications were made (1) where the *DSM-IV* criteria are not applicable to young children (e.g., for CD 5 out of the 15 possible CD criteria [see Table 2] are not measured in the PAPA), (2) evidence from clinical studies of preschoolers supported the use of the RDC-PA developmentally modified *DSM-IV-TR* criteria (i.e., depression, PTSD; Luby et al., 2002b, 2003; Scheeringa et al., 2003), or (3) the high prevalence of certain behaviors in preschoolers indicated a need

TABLE 2
Diagnostic Criteria Used in PAPA Diagnostic Algorithms

Diagnosis	Criteria Used
Depression	RDC-PA (major depressive episode and minor depression); *DSM-IV-TR* (dysthymia)
SAD	*DSM-IV-TR* with inclusion of impairment criterion for a diagnosis
GAD	*DSM-IV-TR*
Specific phobia	*DSM-IV-TR* with inclusion of impairment criterion for a diagnosis
Social phobia	*DSM-IV-TR* with inclusion of impairment criterion for a diagnosis
Selective mutism	*DSM-IV-TR* (impairment criterion included as specified by *DSM-IV-TR*)
PTSD	RDC-PA
ODD	*DSM-IV-TR* using 90th percentile frequency to establish each ODD symptom
CD	*DSM-IV-TR* with 5 CD symptoms excluded (stealing with confrontation, forced sexual activity, breaking into a house or car, running away from home, truancy), but 3-symptom threshold for a diagnosis as defined in *DSM-IV-TR* maintained
ADHD	*DSM-IV-TR* (impairment criterion included as specified by *DSM-IV-TR*)
Enuresis	*DSM-IV-TR* (only applicable to children age ≥5 yr as specified by *DSM-IV-TR*)
Encopresis	*DSM-IV-TR* (only applicable to children age ≥4 yr as specified by *DSM-IV-TR*)
Disability/ impairment	Impairment caused by child's symptoms endorsed in at least one area of functioning in activities or in relationships with others

Note: SAD = separation anxiety disorder; GAD = generalized anxiety disorder; PTSD = post traumatic stress disorder; ODD = oppositional defiant disorder; CD = conduct disorder; ADHD = attention deficit hyperactivity disorder; RDC-PA = Research Diagnostic Criteria-Preschool Age.

to modify the cutpoints for the symptoms. For instance, because the frequency of ODD symptoms such as "often loses temper" is higher in preschoolers than in older children, the ODD algorithm was modified so that each ODD symptom reflected the top 10% of frequency for preschoolers based on PAPA data. Thus, we maintained the 90th percentile frequency cutoff conceptualization of ODD symptomatology of the CAPA (Angold and Costello, 1996) by modifying the criteria frequency levels. A similar approach was taken for the CD symptoms of assaults and lying.

We also required the presence of impairment for diagnosis of three of the anxiety disorders—SAD, specific phobia and social phobia—despite the fact that the *DSM-IV-TR* criteria do not require impairment for a diagnosis (the *DSM-IV-TR* specifies that symptoms must lead to distress or impairment, and all of these symptoms are, by definition in the PAPA, distressing). We made this modification because separation anxieties, social inhibition, and specific fears were relatively common in our subjects (as expected from our knowledge of development), and we were concerned about overdiagnosing these disorders. Because the generalized worry of

GAD is not developmentally normative, we did not include an impairment criterion for this diagnosis. *DSM-IV-TR* diagnoses requiring impairment regardless of distress (i.e., selective mutism, ADHD) include such a criterion in their algorithms.

The PAPA algorithms will be revised over time in response to current and future data about the validity of these diagnostic criteria for preschoolers. The PAPA SAS diagnostic algorithms can be obtained from the first author.

PAPA Training

In this study 6 interviewers conducted the PAPA interviews. All of the interviewers had at least a bachelor's degree. None were mental health specialists. Interviewers had received 1 month of PAPA training, and several had previous experience with the CAPA. Information about training can be found at http://devepi.duhs. duke.edu. The interviews were recorded on audiotape. The interviewer also kept detailed notes throughout the interview, including examples for any positive item, to facilitate coding after completion of the interview. All of the coded interviews were checked before data entry by an interview supervisor with extensive CAPA and PAPA interviewing experience.

Study Design

Subjects were recruited from the Duke (University) Children's Primary Care pediatric clinic. Staffed by both attending and resident pediatricians, the clinic cares for a diverse population of families, drawn not only from the city of Durham but also the surrounding rural areas of Durham County. We wanted to be able to compute statistics that represented unbiased estimates for pediatric clinics. However, we also needed to ensure that we had sufficient numbers of individuals with disorders to allow computation of reasonably stable test-retest reliability statistics for a range of conditions. Hence our choice of a psychopathology screen-stratified design, with oversampling of those with high screen scores. The use of sampling weights permitted unbiased pediatric clinic estimates to be computed from such a stratified sample. An overview of the study design is presented in Figure 1.

So that we could determine whether there were any substantial differences in reliability by gender, age, or race (African American versus non-African American), the sample was also stratified by those factors. We aimed to recruit equal numbers of parents of boys and girls; 2-, 3-, 4-, and 5-year-olds; African Americans and non-African Americans. We also aimed to recruit twice as many screen highs as screen lows. Our target was 6 screen lows and 12 screen highs (total = 18) in each race × age × gender group.

Screening Phase. During the 18 months of data collection, 1,220 parents with children ages 24 to 71 months attended the pediatric clinic during days when we were recruiting subjects. Of these, we missed contacting 29 (2.1%). Thus, 1,191 parents were approached by a screener, who explained the study and sought informed consent for completion of the CBCL/1½–5 (Achenbach and Rescorla, 2000). Children were being seen at the clinic for both well-child and sick-child visits by attending physicians and residents; 98 children were excluded. The exclusion criteria were parent without adequate English to complete the interview (*n* = 48), the index child known (by parent report) to have mental retardation, autism, or other pervasive developmental disorders (*n* = 14: 10 with autism or PDD; 4 with mental retardation without autism or a PDD), a sibling who was already enrolled in the study (*n* = 15), and the adult who brought the child to the clinic could not provide legal consent to participate in the study (*n* = 21). A total of 1,073 parents completed the questionnaire; 118 (9.9%) refused.

Children who obtained a T score ≥55 on the total symptom score of the CBCL (which identifies the top 30% of the general population according to the CBCL/1½–5 norms [Achenbach and Rescorla, 2000]) were considered "screen high." The number of screen highs in our sample was 307 (28.6%). Stratifying by age, gender, and race, we selected 246 (80%) of these screen highs for recruitment, continuing to request participation in the interview phase from members of each age × gender × ethnicity group until their particular cell was full. Of this number, 193 (78%) completed the interview phase. We used a random number generator aimed at selecting 20% of parents whose children had a T score <55 ("screen lows") to take part in the test-retest (TRT) phase of the study. A total of 149 (19.5%) were selected, and of these, 114 (77%) completed both interviews.

TRT Phase

When the parent was selected for recruitment into the TRT phase of the study, the screener arranged a time for an interview. The interviewers were blind to the parent's screen status. PAPA interviews usually took place at the participant's home, but subjects could also choose to be interviewed at Duke University Medical Center. The interview began with the collection of informed consent for the TRT phase. A second interview was scheduled to occur within a period of about 1 week, if possible. The second interview was by an interviewer who was blind to the results of the first.

The overall protocol completion rate from initial contact in the clinic to completion of the second PAPA interview was 70%. There was no significant difference by age, gender, race, or screen status between TRT refusers and subjects. Of the 314 who completed a first interview, 7 (2.2%) refused to complete a second PAPA interview. Parents were asked whether they wanted the results of the CBCL shared with the child's pediatrician, and the results were accordingly transmitted or not. This Duke pediatric clinic was chosen because it has an on-site mental health team to which we could refer parents concerned about their child's mental health.

Data Analysis

Cohen's κ (Cohen, 1960) was used to assess agreement on categorical variables, and the intraclass correlation coefficient (ICC) was used to assess agreement between syndrome scale scores. We computed weighted reliability statistics from the whole sample to produce unbiased estimates of reliability for pediatric primary care. These weights were inversely proportional to the probability of selection into the test-retest sample. We also computed unweighted estimates for the screen high group alone to approximate expected reliability in psychiatric clinic samples.

Differences in κ by gender, age, and ethnicity were tested using a Bayesian approach involving a Markov chain Monte Carlo approach to adjusting for the sampling design (implemented in WinBugs; details are available from A.E.) to produce the correct weighted standard errors. To assess the expected symptom attenuation, the tendency for fewer symptoms to be reported at the second rather than at the first administration (Angold et al., 1996b, 2002; Costello et al., 1984; Jensen et al., 1992, 1995; Lauritsen, 1998; Lucas, 1992; Lucas et al., 1999; Piacentini et al., 1999), we used Poisson regression to reflect the shapes of the scale score distributions and binomial regression for categorical variables.
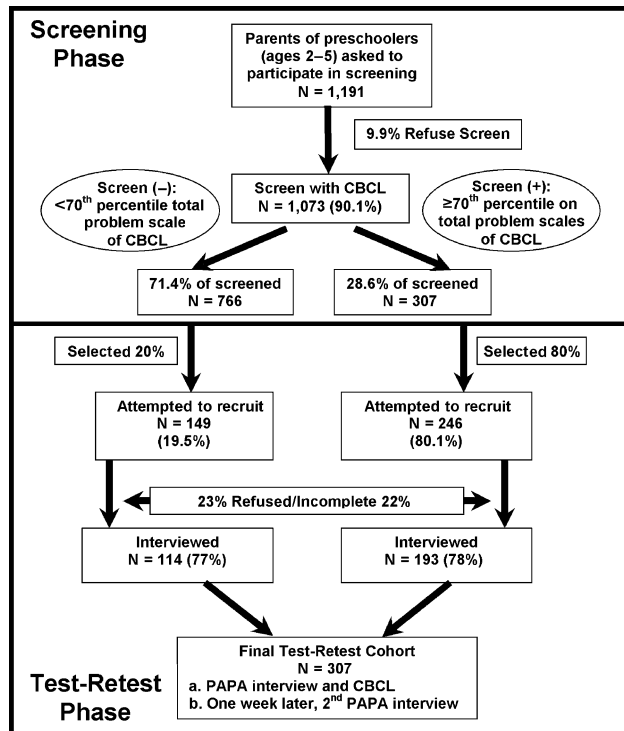
Fig. 1 Design of the PAPA test-retest study (PTRTS).

We conducted weighted analyses with empirical variance estimates using the generalized estimating equations approach implemented in SAS PROC GENMOD (SAS, Cary, NC) to account for the effects of the sampling scheme on both the parameter and variance estimates.

RESULTS

Sample Characteristics

Demographic characteristics of the screened sample, the test-retest sample, and surrounding Durham County where the study was conducted are presented in Table 3.

African American parents were somewhat more likely to participate in the screening (92.5% African Americans agreed versus 87.8% for non-African Americans; $p = 0.007$). There were no significant differences between screen refusers and subjects by gender, age, or Medicaid status.

Of the parents completing the PAPA, 92.7% ($n = 289$) were female. 85.7% ($n = 263$) were biological mothers, 4.9% ($n = 15$) were biological fathers, 1.0% ($n = 3$) were foster mothers, 0.7% ($n = 2$) were unrelated female adults serving as parents, 2.0% ($n = 6$) were adoptive mothers, 0.7% ($n = 2$) were adoptive

fathers, 0.4% ($n = 1$) were live-in male partners of the child's biological mother, 3.3% ($n = 10$) were grandmothers, and 1.6% ($n = 5$) were other female relatives.

Test-Retest Interval

The minimum permitted test-retest interval was 3 days, and the maximum was 1 month. The mean interval was 11 days. The median interval was 7 days (60% of the interviews were conducted between 6 and 8 days after the first interview; 86% were conducted within 14 days of the first interview).

Length of PAPA Administration

Overall, the PAPA took 101 minutes to administer. The first interviews took longer, on average, than the second interviews (107 minutes versus 95 minutes). Interviews with parents whose children met criteria for a *DSM* diagnosis took significantly longer than those about children without a disorder (mean: 127 minutes versus 91 minutes). Significantly increased administration times were also found for impaired children (120 minutes), children with two or more disorders (126 minutes), children with a behavioral disorder, those with an emotional disorder, or those with an SED (all 132 minutes). No significant differences in administration time were noted by age, gender, or race.

Diagnostic Reliability

Table 4 shows levels of agreement for the diagnoses and scale scores generated by the PAPA. For comparison purposes, it also gives reliabilities for the DISC (Shaffer et al., 1999b), the most widely used child psychiatric diagnostic instrument, and the Structured Clinical Interview for DSM (SCID; Williams et al., 1992), a widely used interviewer-based adult instrument.

We found no significant differences in reliability (either κ or ICCs) with respect to age, gender, or race. (Details of this extensive series of analyses are not presented here, but are available from H.L.E.)

Symptom and Diagnostic Attenuation

The left side of Table 5 shows the weighted percentages of individuals meeting diagnostic criteria for a range of individual and combined diagnoses from the first and second interviews. As expected, most diagnoses were made less commonly at the second

TABLE 3
Demographic Characteristics of the PTRTS Subjects Compared
With the Surrounding Community

	PTRTS[a]		Durham County[b]
	Screen	TRT	
Overall N	1,073	307	223,314
Gender, %			
Female	49	46	48
Male	51	54	52
Age, yr (%)			
2	51	30	N/A
3	21	24	
4	26	24	
5	23	22	
Race/ethnicity, %			
AA/black	58	55	40
White/non-Hispanic	32	35	48
Hispanic	2	2	8
Asian	2	1	3
Native American	0.3	0.3	0.3
Other	6	7	6
Medicaid/Medicare, %	43	54	33
Headstart/Early Headstart, %	5	9	4
Family income <$15,000/yr, %	25	31	17
Full-time parental employment, %	63	63	61
Parent education, %			
Some HS	14	9	22
HS graduate	20	30	28
Some college	35	30	27
≥4 yr college	32	31	23

Note: AA = African American; PTRTS = PAPA test-retest study;
HS = high school.
[a] Unweighted percentages.
[b] Information from the 2000 U.S. Census Report (www.factfinder.census.gov).

interview, resulting in an odds ratio (OR) >1, indicating that the odds of having the diagnosis was greater at the first interview. In six cases these differences were significant, and the largest (OR = 1.8) occurred with GAD. In two instances, diagnoses were more commonly made at the second interview, but neither approached statistical significance. However, when all of these effects were combined across diagnoses (the "any disorder" row in Table 5), the OR was only a nonsignificant 1.2. It is also notable that there was no indication of attenuation in reports of disability (OR = 1).

The scale scores shown on the right side of Table 5 present a similar picture, with five areas of significant attenuation, and all but one of the comparisons being in the direction of lower scores at the second interview.

DISCUSSION

Given the well-known decision biases that are inherent in unstructured psychiatric assessments (Achenbach, 1985; Angold, 2002; Tversky and Kahneman, 1974), there is no doubt that the development of standardized structured diagnostic assessments has been a significant advance. Indeed, the use of such assessments has become a necessary requirement for research, and their use has also been strongly advocated for clinical settings as well (Angold, 2002). It is also generally accepted that such instruments should be shown to have acceptable test-retest reliability. However, research on psychopathology in preschoolers has lagged significantly in the development of psychometrically sound structured assessments. Establishing the test-retest reliability of the PAPA is a first step in demonstrating that we can use structured psychiatric interviews, similar to those used with older children and adults, to assess psychiatric symptoms and disorders in preschool children.

The PAPA achieved levels of test-retest reliability similar to those of widely used and well-established measures for older children and adults. Although these were not always as high as one would have liked by any means, whether in preschoolers, older children, or adults, they were high enough to support the continued use of the PAPA in research and clinical diagnosis. The good reliabilities for the assessment of disability and SED suggests that the PAPA will be useful for both researchers and clinicians as they work to define clinically significant psychopathology in preschoolers and to identify the outcomes of early-onset disorders. The fact that the PAPA proved just as reliable when used with the parents of 2-year-olds as with parents of older preschoolers, demonstrates that it is possible to begin assessing psychiatric symptoms and disorders in toddlers, enabling researchers to examine prospectively the early onset of psychopathology. The lack of significant differences in reliabilities for interviews with African American and non-African American parents or with parents of boys and girls suggests that the PAPA can be used with various populations of children.

Across the specific disorders, the reliabilities for the behavioral disorders, particularly ADHD, and depression were better than for the anxiety disorders, with the

TABLE 4
Diagnostic and Scale Score Reliabilities for the PAPA and Selected Comparison Interviews

| DSM Diagnoses | PAPA | | | | Parent DISC[a] | | | SCID[b] | |
| | Full Sample, Weighted (*n* = 307) | | Screen Highs, Unweighted (*n* = 193) | | Community Sample (*n* = 247) | | Clinic Sample (*n* = 84) | Community Sample (*n* = 202) | Clinic Sample (*n* = 390) |
	κ	ICC	κ	ICC	κ	ICC	κ	κ	κ
Depression	0.72	0.71	0.61	0.69	0.55	0.51	0.65	0.47	0.52
SAD	0.60	0.63	0.53	0.68	0.45	0.67	0.58	–	–
GAD	0.39	0.61	0.34	0.61	–	–	0.65	–	–
Specific phobia	0.36	0.57	0.49	0.51	–	0.69	0.96	0.48	0.52
Social phobia	0.54	0.73	0.63	0.75	0.45	0.61	0.54	0.41	0.47
Selective mutism	0.53	–	0.80	–	–	–	–	–	–
PTSD	0.73	0.56	0.79	0.71	–	–	–	–	–
Any anxiety disorder	0.49	0.74	0.50	0.78	–	–	–	–	–
Any emotional disorder	0.54	0.77	0.53	0.80	–	–	–	–	–
ADHD	0.74	0.80	0.67	0.79	0.60	0.61	0.79	–	–
CD	0.60	0.66	0.54	0.70	0.56	0.79	0.43	–	–
ODD	0.57	0.67	0.44	0.62	0.68	0.81	0.54	–	–
Any behavioral disorder	0.53	0.77	0.62	0.80	–	–	–	–	–
Enuresis	0.87	–	0.71	–	–	–	–	–	–
Encopresis	0.74	–	0.64	–	–	–	–	–	–
Any disorder[c]	0.58	0.80	0.56	0.81	–	–	–	0.37	0.61
SED	0.69	–	0.61	–	–	–	–	–	–
Impairment	0.79	0.89	0.71	0.85	–	–	–	–	–

Note: ICC = intraclass correlation; SAD = separation anxiety disorder; GAD = generalized anxiety disorder; PTSD = posttraumatic stress disorder; ADHD = attention-deficit/hyperactivity disorder; CD = conduct disorder; ODD = oppositional defiant disorder; SED = serious emotional disturbance.
[a] Shaffer et al., 1999b.
[b] Williams et al., 1992.
[c] Any disorder excludes elimination disorders.

exception of PTSD and SAD. The lowest reliabilities were found for GAD, specific and social phobias, and selective mutism. There has been remarkably little work on the nosology of anxiety disorders in young children, and these lower reliabilities most likely reflect the need for further research on how to define as well as assess anxiety disorders in young children. These lower reliabilities fall within the range of those reported for anxiety disorders assessed by the DISC and SCID, so there may be a more general problem with the reliability of anxiety assessments.

The good reliability and lack of attenuation for the assessment of disability is encouraging because disability is a key construct for establishing the validity and clinical significance of psychiatric diagnoses in this age group. Further research on disability in preschoolers will have enormous implications not only for our understanding of the course of and prognosis for early-

onset psychiatric syndromes but also for development of effective treatments and interventions for young children and their families.

As has been found for structured interviews for older children (e.g., Angold and Costello, 1995; Jensen et al., 1995; Schwab-Stone et al., 1994), the reliability of the symptom scale scores were, with the exception of the PTSD scale score, better than for the categorical diagnoses. As Shaffer and colleagues (2000) have pointed out, low categorical reliabilities can be misleading because one difference in response can bring the subject above or below the diagnostic threshold. It has also been suggested (Shaffer et al., 1999a) that kappas will tend to be lower in community samples than in clinical samples because more individuals close to the diagnostic cutpoints will be present. However, we did not find significant differences between the weighted reliabilities for the entire

TABLE 5
Prevalence of Diagnoses, Mean Scale Scores, and Attenuation From the First and Second PAPA Interviews

Disorder	Weighted % (Unweighted *n*)				Diagnostic Attenuation, T1/T2 (OR [95% CI] *p* Value)	Mean Scale Scores (SD)		Scale Score Attenuation, T1/T2 (OR [95% CI] *p* Value)
	T1	T2	T1 + T2	T1 or T2		T1	T2	
Depression	2.1 (13)	2.1 (12)	1.5 (8)	2.7 (17)	1.0 (0.7, 1.5) *p* = 1.0	0.8 (1.1)	0.6 (1.0)	**1.3 (1.1, 1.6)** **_p_ = 0.004**
SAD	2.4 (16)	1.6 (10)	1.2 (7)	2.7 (19)	**1.6 (1.0, 2.5)** **_p_ = 0.07**	0.8 (1.0)	0.7 (1.0)	1.2 (0.9, 1.5) *p* = 0.2
GAD	3.8 (27)	2.4 (18)	1.5 (11)	4.7 (34)	**1.8 (1.1, 2.8)** **_p_ = 0.01**	0.6 (1.0)	0.5 (0.9)	**1.3 (1.1, 1.6)** **_p_ = 0.007**
Specific phobia	2.3 (15)	1.7 (10)	0.8 (6)	3.2 (19)	1.3 (0.5, 3.2) *p* = 0.5	0.1 (0.3)	0.2 (0.4)	0.9 (0.6, 1.3) *p* = 0.5
Social phobia	2.1 (14)	2.3 (13)	1.2 (8)	3.2 (20)	0.9 (0.5, 16) *p* = 0.7	0.1 (0.3)	0.1 (0.4)	1.2 (0.9, 1.6) *p* = 0.3
Selective mutism	0.6 (3)	0.4 (3)	0.3 (2)	0.7 (4)	1.5 (0.4, 1.8) *p* = 0.5	-	-	-
PTSD	0.6 (5)	0.7 (5)	0.5 (4)	0.8 (6)	0.8 (0.4, 2.0) *p* = 0.7	1.4 (0.8)	1.4 (0.9)	1.0 (1.0, 1.1) *p* = 0.4
Any anxiety diagnosis	9.4 (55)	6.3 (40)	4.1 (27)	11.6 (68)	**1.6 (1.1, 2.3)** **_p_ = 0.02**	3.1 (2.6)	2.8 (2.4)	**1.1 (1.0, 1.2)** **_p_ = 0.02**
Any emotional diagnosis	10.5 (61)	7.3 (46)	5.3 (33)	12.6 (74)	**1.5 (1.1, 2.1)** **_p_ = 0.02**	3.9 (3.1)	3.4 (3.1)	**1.2 (1.1, 1.3)** **_p_ = 0.004**
ODD	6.6 (38)	7.1 (27)	3.3 (16)	10.4 (49)	0.9 (0.4, 2.0) *p* = 0.8	1.6 (1.5)	1.4 (1.6)	1.2 (1.0, 1.4) *p* = 0.1
CD	3.3 (22)	2.3 (15)	1.8 (11)	3.9 (26)	**1.5 (1.0, 2.2)** **_p_ = 0.06**	0.5 (0.9)	0.4 (0.8)	1.1 (0.8 1.5) *p* = 0.4
ADHD	3.3 (22)	2.9 (20)	2.3 (15)	3.9 (27)	1.1 (0.8, 1.5) *p* = 0.5	1.9 (3.1)	1.6 (2.8)	1.2 (1.0, 1.4) *p* = 0.1
Any behavioral diagnosis	9.0 (56)	10.2 (48)	6.2 (367)	13.0 (68)	0.9 (0.5, 1.5) *p* = 0.6	4.0 (4.8)	3.5 (4.4)	1.2 (1.0, 1.4) *p* = 0.1
Enuresis	15.4 (15)	13.7 (13)	12.9 (12)	16.2 (16)	1.2 (0.9, 1.6) *p* = 0.3	–	–	–
Encopresis	5.1 (8)	4.5 (8)	3.6 (5)	6.0 (11)	1.1 (0.7, 1.8) *p* = 0.6	–	–	–
Any disorder[a]	16.2 (96)	14.1 (74)	9.4 (57)	20.9 (113)	1.2 (0.8, 1.8) *p* = 0.5	7.9 (7.4)	6.9 (6.6)	**1.2 (1.0, 1.3)** **_p_ = 0.02**
SED	12.1 (75)	9.4 (58)	7.8 (47)	13.6 (86)	**1.3 (1.1, 1.6)** **_p_ = 0.007**	–	–	–
Impairment	22.1 (122)	22.8 (122)	18.9 (104)	26.0 (140)	1.0 (0.8, 1.2) *p* = 0.7	2.2 (5.9)	2.1 (5.2)	1.1 (0.9, 1.2) *p* = 0.4

Values in boldface type are statistically significant; OR = odds ratio; CI = confidence interval; SD = standard deviation; SAD = separation anxiety disorder; GAD = generalized anxiety disorder; PTSD = posttraumatic stress disorder; ADHD = attention-deficit/hyperactivity disorder; CD = conduct disorder; ODD = oppositional defiant disorder; SED = serious emotional disorder; T1 = time one (first administration of the PAPA); T2 = time two (second administration of the PAPA).

[a] Excludes elimination disorders.

PAPA test-retest study cohort (approximating a community sample) and the unweighted reliabilities for the screen high groups (approximating a clinical sample).

As seen in all of the studies involving repeated psychiatric assessment (e.g., Angold and Costello, 1995; Jensen et al., 1995; Lucas, 1992; Lucas et al., 1999; Schwab-Stone et al., 1994), attenuation was found for most of the diagnoses and scale scores and reached a level of significance with SAD, GAD, CD, and symptoms of depression and GAD. Further analyses

examining the features of the PAPA items and probes associated with attenuation, as has been done with the DISC (Lucas et al., 1999), will give us the opportunity to understand whether PAPA attenuation reflects systematic bias arising from structural aspects of the PAPA (e.g., complexity of questions, length of the interview) that could be modified to improve reliability.

Clinical Implications

We administered 621 PAPAs (314 first PAPA interviews and 307 repeat PAPAs) during the 18 months of data collection. The fact that only seven parents refused or failed to complete the second PAPA suggests that the experience of being interviewed with the PAPA is acceptable. The administration times, overall and for children with disorders and/or impairment, were comparable to the administration times for psychiatric interviews for older children and adults. The recently developed ePAPA, the electronic version of the PAPA administered on a tablet PC, will decrease administration time, remove the need for separate data entry, and serve as the foundation for the development of a version of the PAPA to be used in clinical practice.

One of the great challenges in studying psychopathology in young children is that we are still working toward the development of a coherent and clinically meaningful psychiatric nosology for this age group. The PAPA was developed to reflect the varied criteria that can be applied to preschool children (e.g., *DSM-IV-TR*, ICD-10, RDC-PA, DC: 0–3R) to make it possible to test empirically the usefulness of the various diagnostic systems they represent across this period of rapid developmental change. The broad coverage provided by the PAPA offers an opportunity to explore the organization of psychiatric syndromes of many types from both categorical and dimensional perspectives. Multiple diagnostic and scalar algorithms can be written that test the validity (and reliability) of various classifications of disorders or empirically derived clusters of symptoms. Examination of the validity of these diagnostic criteria (and, of course, the validity of the interview assessing these criteria) is dependent on having a reliable way to measure the specified criteria. These data demonstrating that we can reliably measure symptoms, disorders, and disability in preschoolers means that it possible to examine the concurrent and predictive validity of various diagnostic nosologies of preschool psychopathology and do the type of psychi-

atric epidemiology with children as young as age 2 that we have done with older children and adults.

Limitations

We employed a pediatric clinic sampling frame, so our results generalize to that setting. Although this is in itself a useful sampling frame because pediatric clinics represent a key clinical contact point for children at this age (few ever make contact with specialist mental health services; e.g., Lavigne et al., 1998), it is possible that we would have seen different levels of reliability had we sampled from other settings. The results of our screening with the CBCL/1½–5 produced a distribution of scores similar to that seen in the general population (71.4% scored at or below the normative general population 70th percentile), so we expect that the weighted analyses for the full sample are a fair reflection of what could be expected in a true random general population sample (as has been found with older children; Costello et al., 1988). We did not use a general population sample because of the enormous expense of community counting and listing for sample identification. Although our sample includes African American children, we had few Hispanic American or Asian American children in our sample, so that our results may not generalize to these populations. This article deals only with the reliability of *DSM-IV-TR* diagnoses and syndromes and the developmentally modified *DSM-IV* criteria presented in the RDC-PA. Syndromes measured in the PAPA but derived from other taxonomies (e.g., DC: 0–3) raise additional definitional and measurement issues that deserve fuller treatment in a separate report.

We see the development of a structured parent interview, in this case the PAPA, as a first step in the development of a comprehensive set of measures for assessing preschool psychopathology. Multiple informants including both parents, other caregivers (teachers, daycare providers, babysitters, other relatives), and the child himself or herself, as well as multiple modes of assessment including structured observational assessments (e.g., the Disruptive Behavior Diagnostic Observational Schedule, a new measure designed to distinguish disruptive from normative behavior in preschool children; Wakschlag et al., 2005), are critical for developing an adequate representation of the child's behaviors and experiences (for a recent review of preschool mental health measures, see Carter et al.,

2004). As reliable and valid measures are developed, we will have to address the critical question of how best to combine the information from different informants and different assessment methods to make diagnostic decisions (Kraemer et al., 2003).

Test-retest reliability is a critical first step in establishing the foundation for examining the validity of these diagnostic categories for preschool children. The validity of an instrument like the PAPA and the validity of the diagnostic system(s) it implements are inextricably intertwined (Robins, 1985), so there can be, even in principle, no simple demonstration of validity. Research on associations among symptoms, disabilities, and risk factors, and studies of the stability of syndromes and diagnoses will ultimately provide the "nomological net" that forms the basis of both measure and construct validity (Cronbach and Meehl, 1955). We hope that this first reliability study will help to open the way for such developments.

REFERENCES

Achenbach TM (1985), *Assessment and Taxonomy of Child and Adolescent Psychopathology*. Beverly Hills, CA: Sage Publications

Achenbach TM, Rescorla LA (2000), *Manual for the ASEBA Preschool Forms and Profiles: An Integrated System of Multi-informant Assessment*. Burlington, VT: University of Vermont Department of Psychiatry

American Psychiatric Association (2000), *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision*. Washington, DC: American Psychiatric Press

Angold A (2002), Diagnostic interviews with parents and children. In: *Child and Adolescent Psychiatry: Modern Approaches, Fourth Edition*, Rutter M, Taylor E, eds. Oxford: Blackwell Scientific, pp 32–51

Angold A, Costello EJ (2000a), *A Review of the Issues Relevant to the Creation of a Measure of Disability in Children Based on the World Health Organization's International Classification of Functioning and Disability (ICIDH-2)*. Geneva: World Health Organization

Angold A, Costello EJ (1995), A test-retest reliability study of child-reported psychiatric symptoms and diagnoses using the Child and Adolescent Psychiatric Assessment (CAPA-C). *Psychol Med* 25:755–762

Angold A, Costello EJ (1996), Toward establishing an empirical basis for the diagnosis of oppositional defiant disorder. *J Am Acad Child Adolesc Psychiatry* 35:1205–1212

Angold A, Costello EJ (2000), The Child and Adolescent Psychiatric Assessment (CAPA). *J Am Acad Child Adolesc Psychiatry* 39:39–48

Angold A, Egger HL (2004), Psychiatric diagnosis in preschool children. In: *Handbook of Infant, Toddler, and Preschool Mental Health Assessment*, DelCarmen-Wiggins R, Carter A, eds. New York: Oxford University Press, pp 123–139

Angold A, Erkanli A, Costello EJ, Rutter M (1996a), Precision, reliability and accuracy in the dating of symptom onsets in child and adolescent psychopathology. *J Child Psychol Psychiatry* 37:657–664

Angold A, Erkanli A, Loeber R, Costello EJ, Van Kammen W, Stouthamer-Loeber M (1996b), Disappearing depression in a population sample of boys. *J Emotional Behav Disord* 4:95–104

Angold A, Erkanli A, Silberg J, Eaves L, Costello E (2002), Depression scale scores in 8–17-year-olds: Effects of age and gender. *J Child Psychol Psychiatry* 43:1052–1063

Angold A, Prendergast M, Cox A, Harrington R, Simonoff E, Rutter M (1995), The Child and Adolescent Psychiatric Assessment (CAPA). *Psychol Med* 25:739–753

Boris NW, Zeanah CH, Larrieu JA, Scheeringa MS, Heller SS (1998), Attachment disorders in infancy and early childhood: a preliminary investigation of diagnostic criteria. *Am J Psychiatry* 155:295–297

Briggs-Gowan MJ, Horwitz SM, Schwab-Stone ME, Leventhal JM, Leaf PJ (2000), Mental health in pediatric settings: distribution of disorders and factors related to service use. *J Am Acad Child Adolesc Psychiatry* 39:841–849

Carter AS, Briggs-Gowan MJ, Davis NO (2004), Assessment of young children's social-emotional development and psychopathology: recent advances and recommendations for practice. *J Child Psychol Psychiatry* 45:109–134

Cohen J (1960), A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46

Cordeiro M, Caldeira da Silva P, Goldschmidt T (2003), Diagnostic classification: results from a clinical experience of three years with DC: 0–3. *Infant Ment Health J* 24:349–364

Costello AJ, Edelbrock CS, Dulcan MK, Kalas R, Klaric SH (1984), *Development and Testing of the NIMH Diagnostic Interview Schedule for Children in a Clinic Population: Final Report (contract no. RFP-DB-81-0027)*. Rockville, MD: NIMH Center for Epidemiologic Studies

Costello EJ, Costello AJ, Edelbrock C, Burns BJ, Dulcan MK, Brent D, Janiszewski S (1988), Psychiatric disorders in pediatric primary care: prevalence and risk factors. *Arch Gen Psychiatry* 45:1107–1116

Cronbach LJ, Meehl PE (1955), Construct validity in psychological tests. *Psychol Bull* 52:281–302

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

Egger HL, Angold A (2004), The Preschool Age Psychiatric Assessment (PAPA): a structured parent interview for diagnosing psychiatric disorders in preschool children. In: *Handbook of Infant, Toddler, and Preschool Mental Assessment*, DelCarmen-Wiggins R, Carter A, eds. New York: Oxford University Press, pp 223–243

Egger HL, Ascher BH, Angold A (1999), *The Preschool Age Psychiatric Assessment: Version 1.1. Unpublished Interview Schedule*. Durham, NC: Center for Developmental Epidemiology, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center

Federal Register (1993), 58:29425

Guedeney N, Guedeney A, Rabouam C, Mintz A-S, Danon G, Huet M, Jacqueman F (2003), The Zero-to-Three diagnostic classification: a contribution to the validation of this classification from a sample of 85 under-threes. *Infant Ment Health J* 24:313–336

Jensen P, Roper M, Fisher PW, Piacentini J, Canino G, Richters J, Rubio-Stipec M, Dulcan M, Goodman S, Davies M, Rae D, Shaffer D, Bird H, Lahey B, Schwab-Stone M (1995), Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1): parent, child, and combined algorithms. *Arch Gen Psychiatry* 52:61–71

Jensen PS, Shaffer D, Rae D, Canino G, Bird HR (1992), Attenuation of the Diagnostic Interview Schedule for Children (Disc 2.1): Sex, Age and IQ Relationships. Paper presented at the 39th Annual Meeting of the American Academy of Child and Adolescent Psychiatry, Washington, DC, October.

Kashani JH, Allan WD, Beck NCJ, Bledsoe Y, Reid JC (1997), Dysthymic disorder in clinically referred preschool children. *J Am Acad Child Adolesc Psychiatry* 36:1426–1433

Kashani JH, Holcomb WR, Orvaschel H (1986), Depression and depressive symptoms in preschool children from the general population. *Am J Psychiatry* 143:1138–1143

Kashani JH, Ray JS, Carlson GA (1984), Depression and depressive-like states in preschool-age children in a child development unit. *Am J Psychiatry* 141:1397–1402

Keenan K, Shaw DS, Walsh B, Delliquadri E, Giovannelli J (1997), DSM-III-R disorders in preschool children from low-income families. *J Am Acad Child Adolesc Psychiatry* 36:620–627

Keenan K, Wakschlag LS (2000), More than the terrible twos: the nature and severity of behavior problems in clinic-referred preschool children. *J Abnorm Child Psychol* 28:33–46

Keenan K, Wakschlag LS (2002), Can a valid diagnosis of disruptive behavior disorder be made in preschool children? *Am J Psychiatry* 159:351–358

Keren M, Feldman R, Tyano S (2003), A five-year Israeli experience with the DC: 0–3 classification system. *Infant Ment Health J* 24:337–348

Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ (2003), A new approach to integrating data from multiple informants in psychiatric assessment and research: mixing and matching contexts and perspectives. *Am J Psychiatry* 160:1566–1577

Lahey BB, Pelham WE, Stein MA, Loney J, Trapani C, Nugent K, Kipp H, Schmidt E, Lee S, Cale M, Gold E, Hartung CM, Willcutt E, Baumann B (1998), Validity of DSM-IV attention-deficit/hyperactivity disorder for younger children. *J Am Acad Child Adolesc Psychiatry* 37:695–702

Lauritsen JL (1998), The age-crime debate: assessing the limits of longitudinal self-report data. *Soc Forces* 77:127–154

Lavigne JV, Arend R, Rosenbaum D, Binns HJ, Christoffel KK, Burns A, Smith A (1998), Mental health service use among young children receiving pediatric primary care. *J Am Acad Child Adolesc Psychiatry* 37:1175–1183

Lavigne JV, Gibbons RD, Christoffel KK, Arend R, Rosenbaum D, Binns H, Dawson N, Sobel H, Issacs C (1996), Prevalence rates and correlates of psychiatric disorders among preschool children. *J Am Acad Child Adolesc Psychiatry* 35:204–214

Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E (1989), Autism diagnostic observation schedule. A standardized observation of communicative and social behavior. *J Autism Dev Disord* 19:185–212

Lord C, Rutter M, LeCouteur A (1994), Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 24:659–685

Luby J, Heffelfinger A, Measelle JR, Ablow JC, Essex M, Dierker L, Harrington R, Kraemer H, Kupfer DJ (2002a), Differential performance of the MacArthur HBQ and DISC-IV in identifying DSM-IV internalizing psychopathology in young children. *J Am Acad Child Adolesc Psychiatry* 41:458–466

Luby J, Heffelfinger A, Mrakotsky C, Brown K, Hessler M, Wallis J, Spitznagel E (2003), The clinical picture of depression in preschool children. *J Am Acad Child Adolesc Psychiatry* 42:340–348

Luby J, Heffelfinger A, Mrakotsky C, Hessler M, Brown K, Hildebrand T (2002b), Preschool major depressive disorder: preliminary validation for developmentally modified DSM-IV criteria. *J Am Acad Child Adolesc Psychiatry* 41:928–937

Lucas CP (1992), The order effect: reflections on the validity of multiple test presentations. *Psychol Med* 22:197–202

Lucas CP, Fisher P, Piacentini J, Zhang H, Jensen PS, Shaffer D, Dulcan M, Schwab-Stone M, Regier D, Canino G (1999), Features of interview questions associated with attenuation of symptom reports. *J Abnorm Child Psychol* 27:429–437

Piacentini J, Roper M, Jensen P, Lucas C, Fisher P, Bird H, Bourdon K, Schwab-Stone M, Rubio-Stipec M, Davies M, Dulcan M (1999), Informant-based determinants of symptom attenuation in structured child psychiatric interviews. *J Abnorm Child Psychol* 27:417–428

Reams R (1999), Children birth to three entering the state's custody. *Infant Ment Health J* 20:166–174

Robins LN (1985), Epidemiology: reflections on testing the validity of psychiatric interviews. *Arch Gen Psychiatry* 42:918–924

Rothbart MK, Ellis LK, Rueda MR, Posner MI (2003), Developing mechanisms of temperamental effortful control. *J Pers* 71:1113–1143

Scheeringa MS, Zeanah C, Myers L, Putnam F (2003), New findings on alternative criteria for PTSD in preschool children. *J Am Acad Child Adolesc Psychiatry* 42:561–571

Scheeringa MS, Zeanah CH, Drell MJ, Larrieu JA (1995), Two approaches to the diagnosis of posttraumatic stress disorder in infancy and early childhood. *J Am Acad Child Adolesc Psychiatry* 34:191–200

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

Shaffer D, Fisher P, Lucas CP, Dulcan MK, Schwab-Stone ME (2000), NIMH diagnostic interview schedule for children version IV (NIMH DISC-IV): description, differences from previous versions, and reliability of some common diagnoses. *J Am Acad Child Adolesc Psychiatry* 39: 28–38

Shaffer D, Fisher PW, Lucas CP (1999a), Respondent-based interviews. In: *Diagnostic Assessment in Child and Adolescent Psychopathology*, Shaffer D, Lucas CP, Richters JE, eds. New York: The Guilford Press, pp 3–33

Shaffer D, Lucas CP, Richters JE (1999b), *Diagnostic Assessment in Child and Adolescent Psychopathology*. New York: The Guilford Press

Shaw DS, Owens EB, Giovannelli J, Winslow EB (2001), Infant and toddler pathways leading to early externalizing disorders. *J Am Acad Child Adolesc Psychiatry* 40:44–51

Shaw DS, Owens EB, Vondra JI, Keenan K, Winslow EB (1996), Early risk factors and pathways in the development of early disruptive behavior problems. *Dev Psychopathol* 8:679–699

Speltz M, DeKlyen M, Greenberg M, Dryden M (1995), Clinic referral for oppositional defiant disorder: relative significance of attachment and behavioral variables. *J Abnorm Child Psychol* 23:487–507

Speltz M, McClellan J, DeKlyen M, Jones K (1999), Preschool boys with oppositional defiant disorder: clinical presentation and diagnostic change. *J Am Acad Child Adolesc Psychiatry* 38:838–845

Task Force on Research Diagnostic Criteria: Infancy and Preschool (2003), Research diagnostic criteria for infants and preschool children: the process and empirical support. *J Am Acad Child Adolesc Psychiatry* 42: 1504–1512

Thomas JM, Clark R (1998), Disruptive behavior in the very young child: diagnostic classification: 0–3 guides identification of risk factors and relational interventions. *Infant Ment Health J* 19:229–244

Tversky A, Kahneman D (1974), Judgement under uncertainty: heuristics and biases. *Science* 185:1124–1131

Volkmar FR, Lord C, Bailey A (2004), Autism and pervasive developmental disorders. *J Child Psychol Psychiatry* 45:135–170

Wakschlag LS, Leventhal B, Briggs-Gowan MJ, Danis B, Keenan K, Hill C, Egger HL, Cicchetti D, Carter A (2005), Defining the "disruptive" in preschool behavior: what diagnostic observation can teach us. *Clin Child Fam Psychol Rev* 8:183–201

Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HG, Rounsaville B (1992), The Structured Clinical Interview for DSM-III-R (SCID): II. Multisite test-retest reliability. *Arch Gen Psychiatry* 49:630–636

World Health Organization (1993), *ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: World Health Organization

World Health Organization (2001), *ICF: International Classification of Functioning, Disability and Health*. Geneva: World Health Organization

Zero to Three (1994), *Diagnostic Classification: 0–3: Diagnostic Classification of Mental Health and Developmental Disorders of Infancy and Early Childhood*. Arlington, VA: National Center for Clinical Infant Programs

Zero to Three (2005), *Diagnostic Classification: 0–3R: Diagnostic Classification of Mental Health and Developmental Disorders of Infancy and Early Childhood, Revised Edition*. Washington, DC: Zero to Three Press