Criterion Validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3)

MARY E. SCHWAB-STONE, M.D., DAVID SHAFFER, M.D., MINA K. DULCAN, M.D., PETER S. JENSEN, M.D., PRUDENCE FISHER, M.S., HECTOR R. BIRD, M.D., SHERRYL H. GOODMAN, PH.D., BENJAMIN B. LAHEY, PH.D., JUDITH H. LICHTMAN, M.P.H., GLORISA CANINO, PH.D., MARITZA RUBIO-STIPEC, M.A., AND DONALD S. RAE, M.A.

ABSTRACT

Objective: To examine the criterion validity of the NIMH Diagnostic Interview Schedule for Children (DISC) Version 2.3 in the NIMH Methods for the Epidemiology of Child and Adolescent Mental Disorders (MECA) Study, using a design that permitted several comparisons of DISC-generated diagnoses with diagnoses based on clinician symptom ratings. **Method:** Two hundred forty-seven youths were selected from the 1,285 parent–youth pairs that constituted the four-site MECA sample. Subjects who screened positive for any of the five diagnostic areas under investigation in the validity study (attention-deficit hyperactivity disorder, oppositional defiant disorder, conduct disorder, depressive disorder, and the major anxiety disorders) were recruited, as well as a comparable number of screen negatives. Clinicians reinterviewed separately both the youth and the primary caregiver using the DISC followed by a clinical-style interview, and then they rated the presence of symptoms and impairment. Computer algorithms combined this information into diagnoses using comparable rules for both DISC and clinical rating diagnoses. **Results:** In general, the DISC showed moderate to good validity across a number of diagnoses. **Conclusions:** Results suggest some specific diagnostic areas in which further revision of the DISC is warranted. Three main sources of variability in DISC–clinician diagnostic agreement were evident over and above that due to the instrument itself, including (1) the informant used, (2) the algorithm applied in synthesizing symptom reports, and (3) the design of the validity comparison. *J. Am. Acad. Child Adolesc. Psychiatry,* 1996, 35(7):878–888. **Key Words:** Diagnostic Interview Schedule for Children, diagnosis, validity, epidemiology, assessment.

In order for the results of community-based studies in psychiatry to be credible to clinicians, researchers, and policymakers, the diagnostic assessment procedures that they use must be shown to generate clinically meaningful diagnoses. In child psychiatry, considerable effort has been devoted to developing instruments suitable for use in community studies and to demonstrating their diagnostic validity. This article reports the results of a validation study of the National Institute of Mental Health (NIMH) Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3) that was conducted in the context of a multisite methodological research effort, the NIMH Methods for the Epidemiology of Child and Adolescent Mental Disorders (MECA) Study (Lahey et al., 1996).

This is the most recent of several validity studies that have been undertaken in the course of writing and revising the DISC. In the initial methodological study conducted on a patient sample in Pittsburgh (Costello et al., 1984), agreement between DISC-1 diagnoses (*DSM-III*) and those generated in clinical case conferences was generally poor. Cohen et al. (1987) compared diagnoses generated by the DISC-1 with those derived from the clinician-administered Schedule for Affective Disorders and Schizophrenia for School-Age Children for a community sample of children

Accepted November 7, 1995.

From the Yale Child Study Center, New Haven, CT (Dr. Schwab-Stone and Ms. Lichtman); Columbia University, New York (Drs. Shaffer, Bird, and Ms. Fisher); Children's Memorial Hospital and Northwestern University Medical School, Chicago (Dr. Dulcan); the NIMH, Rockville, MD (Drs. Jensen and Rae); Emory University, Atlanta (Dr. Goodman); University of Chicago (Dr. Lahey); and the University of Puerto Rico, San Juan (Dr. Canino and Ms. Rubio-Stipec).

The authors gratefully acknowledge the extensive efforts of Beverly Crowther, M.Ed., in preparing the manuscript and of Louis P. Florio, M.S., in conducting the analyses.

Reprint requests to Dr. Schwab-Stone, Yale Child Study Center, P.O. Box 207900, 230 S. Frontage Road, New Haven, CT 06520-7900.

and adolescents and found only moderate agreement. Piacentini et al. (1993) examined the validity of a revised version of the instrument (DISC-R) by comparing DISC-R diagnoses (*DSM-III-R*) with those obtained from a semistructured clinical interview for a sample of outpatients. Overall, agreement between the DISC-R and the clinical method was moderate; however, the clinical validation procedure, i.e., the "standard," was found to be less reliable than the DISC-R, thus limiting interpretations about the validity of the DISC-R. Fisher et al. (1993) examined the sensitivity of the DISC version 2.1 (*DSM-III-R*) for certain rare disorders, such as eating disorders, major depressive disorder, and obsessive-compulsive disorder. Since community rates for these disorders are low, it is difficult to obtain sufficient numbers in community samples to determine the validity of the DISC. Good to excellent sensitivity was found for DISC-generated diagnoses when compared with the clinical diagnoses of the specialty diagnostic and treatment centers from which the subjects were recruited.

In several ways the current study builds on these and other previous efforts to examine the validity of structured diagnostic interviews. First, most validity studies have utilized clinic patients in order to generate symptom-rich samples for evaluation. Because the DISC is intended for use in community-based studies, the proper evaluation of its performance requires a representative community sample, at least for disorders for which sufficient cases can be obtained from community samples. Accurate case identification for nonreferred individuals is critical to health care planning and prevention efforts, as well as to studies on risk factors and etiology.

Second, previous validity studies of structured interviews for children have often compared interview-generated diagnoses to global diagnostic ratings made by clinicians. Such a design may be an unfair test of the instrument, however, since clinicians' ratings may draw on broader information bases than the structured assessments, i.e., best-estimate and chart review procedures allow information from a variety of sources (Piacentini et al., 1993). The current study used the same information sources for both the DISC and clinician interviews and required sequential clinician ratings to reflect the clinical processing of information from the most molecular (question about a symptom) to the most global level (diagnosis).

Third, in studies comparing the DISC and a clinical interview method, it has usually not been possible to examine test-retest and criterion validity in the same study. The study by Piacentini et al. (1993) was designed to allow both types of methodological assessment and found that lack of test-retest reliability plagued the clinical assessment to an even greater degree than the DISC. This has suggested the importance of focusing the current study on the issue of whether the DISC and clinicians *elicit* comparable symptom information. The design for this validation study of the DISC makes it possible to examine criterion validity when there is a retest interval between the DISC and the clinical assessments and also when there is virtually no interval, i.e., concurrent criterion validity.

A clinical validation procedure was developed for this study of 247 subjects drawn from the four sites of the MECA project. This article presents the design of that procedure and the results from that study.

METHOD

Subjects

The sample was constructed as a subsample of the 1,285 parent–youth pairs who were interviewed as described by Lahey et al. (1996). After the survey interview, which included the laptop computer version of the DISC (PC-DISC) administered to the youth and primary caregiver, diagnostic algorithms were run to determine whether the subject met *DSM-III-R* criteria for one or more of the diagnoses under consideration in the validation study. From this screening procedure either a "screen positive" or "screen negative" designation was made, with screen positive status denoting a DISC diagnosis on either the parent or child interview in any of the five diagnostic categories: attention-deficit hyperactivity disorder (ADHD); oppositional defiant disorder (ODD); conduct disorder (CD); depressive disorder (major depressive disorder or dysthymia considered as a single category, MDD/Dys); and the major anxiety disorders considered as a single category that included overanxious disorder (OAD), separation anxiety disorder (SAD), social phobia (SoPh), generalized anxiety disorder, agoraphobic disorder, and panic disorder. Depressive and anxiety disorders were considered in broader categories because it was not expected that sufficient cases would be identified to allow examination at the level of the constituent individual diagnoses. Screen negatives had no parent or child DISC diagnosis in any of the five diagnostic areas. For the validation study, each of the four sites (Columbia, Emory, University of Puerto Rico, and Yale) followed standard procedures to select a minimum of five parent–child subject pairs for each disorder category and a corresponding number of screen negative candidates. From this pool, eligible families were recruited for the validation interviews. The response rate for the validation reinterview study was 88.4% (screen positive = 134, screen negative = 113).

Table 1 shows the demographic characteristics of the validation sample, which are comparable with those of the larger MECA

TABLE 1
Demographic Characteristics of Validation Sample (N = 247)

	Georgia		New Haven		New York		Puerto Rico		Overall	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Male	28	57.1	38	65.5	23	44.2	41	46.6	130	52.6
Female	21	42.9	20	34.5	29	55.8	47	53.4	117	47.4
Age										
9–11 yr	17	34.7	15	25.9	19	36.5	29	33.0	80	32.4
12–14 yr	18	36.7	27	46.6	12	23.1	36	40.9	93	37.7
15–18 yr	14	28.6	16	27.6	21	40.4	23	26.1	74	30.0
Ethnicity										
White (non-Hispanic)	34	69.4	51	87.9	37	71.2	0	0.0	122	49.4
African-American	12	24.5	3	5.2	8	15.4	0	0.0	23	9.3
Hispanic	1	2.0	1	1.7	4	7.7	88	100.0	94	38.1
Other	2	4.1	3	5.2	3	5.8	0	0.0	8	3.2
Annual household income										
<$10,000	3	6.1	4	6.9	3	5.8	47	54.0	57	23.2
$10,000–$24,000	12	24.5	5	8.6	9	17.3	24	27.6	50	20.3
$25,000–$64,000	25	51.0	32	55.2	14	26.9	14	16.1	85	34.6
$65,000–$99,000	9	18.4	17	29.3	13	25.0	1	1.1	40	16.3
>$100,000	0	0.0	0	0.0	13	25.0	1	1.1	14	5.7

sample (see Lahey et al., 1996). Fifty-three percent of the validation study subjects were male and 47% were female. There was a fairly uniform distribution across the age groups. (Several children who were 17 years old at the time of initial interview turned 18 by the time of the second interview; hence the inclusion of a few 18-year-olds in the oldest age group.) Approximately half (49%) of the respondents were Caucasian, 38% were Hispanic, 9% were African-American, and 3% were of other ethnic origin. The distribution of annual household income varied by geographic site and resembles the distribution previously presented for the overall MECA sample (Lahey et al., 1996).

Procedure

The procedure for the validation study was developed to allow multiple diagnostic comparisons, including comparison of DISC and clinician assessments both with and without a retest interval. Procedures for the initial survey interview have been described in detail by Lahey et al. (1996). For the validation study, parent–child pairs who were selected and who agreed to participate were reinterviewed separately by interviewers blind to the previous assessment. Eighty-three percent of the reinterviews were conducted within 1 to 15 days of the initial lay interview; the remainder were completed later.

Clinician interviewers were used for the DISC reinterview and subsequent clinical-style interview. The clinicians were mental health professionals (at least master's degree level) who were judged by leaders at each site to have good clinical skills, knowledge of *DSM-III-R*, and facility with phenomenological psychiatric diagnosis. At the reinterview these clinicians administered an abbreviated version of the DISC-2.3 that included the five major diagnostic areas evaluated in the study. They were trained to administer the interview in the same standard, structured manner as the lay interviewers, using the laptop PC-DISC computer version. As in the initial DISC assessment (Lahey et al., 1996), the Spanish version of the DISC (Bravo et al., 1993; Ribera et al., 1996) was administered at the Puerto Rico site. During the DISC interview, the clinician was able to flag questions (using a special keystroke on the computer) that he or she sensed might have been misinterpreted by the subject or answered differently under another style of questioning. Upon completion of the structured interview, the respondent took a break while the clinician marked on a standard form (available from author on request) the question numbers for all positive responses and for those questions flagged because of some clinical doubt about their veracity. The clinician then conducted a clinical-style interview with the respondent in which the goals were (1) to assess whether items to which the subject had responded positively actually represented clinically meaningful symptoms, and (2) to resolve the clinician's doubt about any responses flagged during the interview because of uncertainty about their clinical meaning. Positive and doubtful negative responses were selected for reevaluation because repeating all questions was not feasible, there were past concerns about false-positives on the DISC, and there was concern that probing all negative responses might seem like badgering to the subjects. Clinicians made ratings of the presence and duration of symptoms, criteria, and diagnoses on the standard form. When any symptoms were present, clinicians also rated level of impairment and the contexts (home, school, peers) in which impairment was manifest. CGAS scores (Shaffer et al., 1983) were assigned for all subjects.

Diagnoses based on the clinician's DISC interview were generated with the same scoring algorithms and in the same manner as in the larger survey (see Shaffer et al., 1996). Computerized scoring algorithms were written to generate comparably constructed diagnoses based on the clinician's symptom ratings. These algorithms are virtually identical with those for the DISC, since the rating form was tailored to make direct comparison possible.

Analytic Strategy

For the five diagnostic categories (ADHD, ODD, CD, depressive disorder, any anxiety disorder), three basic comparisons that bear

on the validity of the DISC have been examined: (1) test-retest agreement for lay and clinician-administered DISC, (2) agreement between lay DISC diagnoses and diagnoses generated from clinician symptom ratings generated after the clinical-style interview, and (3) agreement between diagnoses from the clinician-administered DISC and diagnoses generated from the clinician symptom ratings after the clinical-style interview. These comparisons are shown in schematic form in Figure 1. Diagnostic agreement was evaluated using the κ statistic (Cohen, 1960). Standard errors for κ were calculated according to the method of Fleiss (1981). In the results reported below, a minimum of five cases for each diagnosis were required from either the initial lay DISC or from the subsequent assessment for examination of agreement using κ. Although it was intended that the validity of the depressive and anxiety disorders would be examined at the group rather than the specific diagnostic level, for MDD, Dys, OAD, SAD, and SoPh, there were sufficient numbers across most comparisons to permit examination as specific diagnoses.

Comparison A. Test-Retest Agreement: Lay DISC Diagnoses versus Clinician-Administered DISC Diagnoses. Since retest agreement sets a limit on the agreement that can be expected between the initial lay-administered DISC diagnoses and the diagnoses derived from clinician ratings, test-retest effects were examined first. In this comparison (A on Fig. 1), diagnoses generated from the standard lay-administered DISC were compared with those obtained from the clinician-administered DISC (retest) in which the clinicians followed structured procedures identical with those used by the lay interviewers. In principle, no clinical effect enters this comparison since the rules for DISC administration were highly specified. Thus, the first analysis (comparison A) focuses on test-retest agreement between the two assessments, without any assumption that the clinician-administered DISC is a standard with respect to diagnosis.

Comparison B. DISC Validity with Retest: Lay DISC Diagnoses versus Diagnoses Generated from Clinician Symptom Ratings. The second analysis compares lay-administered DISC diagnoses and those derived from clinician symptom ratings made on the basis of information obtained from the semistructured clinical-style interview. This comparison addresses the validity of the DISC procedure for eliciting symptom information. Comparable algorithms are used to combine symptom reports into diagnoses for the lay DISC and for the clinician symptom ratings; thus there is minimal variability between the two methods of synthesizing diagnoses once symptoms have been assessed. As noted above, because of the interval between the lay and clinician administrations, this comparison builds in a decrement in agreement based on comparison to a retest interview (Jensen et al., 1995; Robins, 1985). Lack of agreement in this analysis is presumed to reflect *both* this retest effect, as well as the difference between the lay-administered DISC and the clinical interview as tools for eliciting information about psychiatric symptoms.

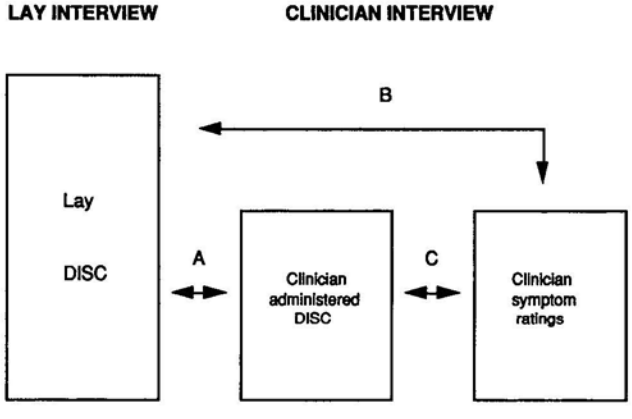

Fig. 1 Analytic Plan. Diagnostic comparisons: *A.* Lay DISC diagnoses versus clinician-administered DISC diagnoses. *B.* Lay DISC diagnoses versus diagnoses from clinician symptom ratings. *C.* Clinician-administered DISC diagnoses versus diagnoses from clinician symptom ratings. DISC = Diagnostic Interview Schedule for Children.

Comparison C. DISC Concurrent Validity: Clinician-Administered DISC Diagnoses versus Diagnoses Generated from Clinician Symptom Ratings. The third analysis bypasses the retest effect by comparing diagnoses from the clinician-administered DISC (DISC information ascertained in the standard structured manner) and diagnoses generated from clinician symptom ratings made after the clinical interview. Like the preceding comparison, this analysis examines the validity of the DISC with respect to its capacity to elicit symptoms accurately; however, it eliminates the retest interval by comparing assessments made on the basis of information gathered at sequential stages of one extended interview. Thus, this comparison aims to evaluate the concurrent validity of the DISC for eliciting symptom information.

Diagnostic Algorithms. For each comparison, it is possible to apply a number of different diagnostic algorithms, depending on the definition of diagnosis that is of interest. Thus for each of the three comparisons (A, B, and C above), three different algorithms were applied: (1) "criteria only"—diagnoses based purely on the *DSM-III-R*–specified symptom counts and their required durations and onsets (e.g., for ADHD, 8 of 14 symptoms and 6-month duration of disturbance with onset before age 7 years); (2) "criteria & impairment"—diagnoses based on *DSM-III-R*–specified criteria with the additional requirement of impairment from the symptoms of that disorder (from DISC questions that ask whether the constellation of reported symptoms caused problems at home, with peers, or at school/work); and (3) "criteria, impairment & CGAS ≤70"—diagnoses constructed to include not only impairment from the disorder being examined but also the requirement of a CGAS score of 70 or less, indicating a compromised level of overall functioning sufficient to be considered a case (Shaffer et al., 1983). Before comparisons were made at this level of the analysis, the lay interviewer CGAS score (Bird et al., unpublished) was applied as a cutoff (≤70) to the original lay-administered DISC "criteria & impairment" diagnoses, while a clinician CGAS rating was applied to the clinician-administered DISC "criteria & impairment" diagnoses. In comparison C, a set of sequential symptom ratings by the clinician is cut by one CGAS rating made by the clinician after the clinical-style interview. Here the CGAS can be viewed as restricting the pool to a more severely disturbed group for which the "criteria & impairment" diagnoses from the DISC and from the semistructured interview are compared, i.e., the CGAS rating is not independent across the two assessments being compared, as it is in comparisons A and B.

In addition to having different algorithms based on varying requirements about the presence of impairment, two types of algorithm are possible depending on whether reports from one or from both informants are used. Single-informant algorithms are identical for parent and for youth forms of the DISC. A "combined" algorithm integrates criteria reported by both informants; thus a criterion is considered to be present if reported by either the parent or the youth, and a diagnosis is made if the total, summed across both informants, meets the *DSM-III-R* rule for the diagnosis (Shaffer et al., 1996). In the comparisons that use the CGAS cut,

TABLE 2
Comparison A: Test-Retest Agreement of Lay DISC Diagnoses versus Clinician-Administered DISC Diagnoses

	Criteria Only		Criteria & Impairment		Criteria, Impairment & CGAS ≤70	
Diagnosis	κ	SE	κ	SE	κ	SE
Parent informant						
ADHD	.60	.064	.65	.064	.80	.064
ODD	.68	.064	.65	.064	.73	.064
CD	.56	.064	.56	.064	.59	.064
Depressive	.50	.062	.50	.062	.56	.060
MDD	.55	.063	.54	.063	.62	.061
Dys	.30	.063	.34	.062	.38	.060
Anxiety	.56	.064	.52	.064	.64	.063
OAD	.60	.062	.51	.063	.76	.063
SAD	.45	.064	.41	.064	.39	.062
SoPh	.45	.064	.20	.064	.35	.063
Youth informant						
ADHD	.10	.055	.19	.038	.22	.040
ODD	.18	.051	.20	.053	.22	.048
CD	.64	.063	.43	.061	.48	.060
Depressive	.35	.058	.29	.057	.38	.055
MDD	.37	.056	.34	.060	.45	.058
Dys	.43	.060	.44	.053	.44	.053
Anxiety	.39	.057	.19	.051	.19	.052
OAD	.28	.050	.10	.041	.19	.038
SAD	.27	.063	.15	.060	—	—
SoPh	.33	.058	.08	.056	.14	.054
Combined parent & youth						
ADHD	.48	.063	.56	.063	.64	.063
ODD	.59	.062	.56	.062	.61	.061
CD	.66	.063	.55	.063	.58	.062
Depressive	.45	.060	.37	.061	.43	.059
MDD	.48	.059	.41	.061	.48	.059
Dys	.35	.062	.35	.062	.35	.060
Anxiety	.47	.062	.43	.061	.54	.061
OAD	.52	.060	.48	.059	.60	.060
SAD	.49	.064	.45	.063	.55	.063
SoPh	.44	.062	.26	.063	.43	.061

Note: DISC = Diagnostic Interview Schedule for Children; ADHD = attention-deficit hyperactivity disorder; ODD = oppositional defiant disorder; CD = conduct disorder; MDD = major depressive disorder; Dys = dysthymia; OAD = overanxious disorder; SAD = separation anxiety disorder; SoPh = social phobia; Depressive = MDD or Dys; Anxiety = any of the following: social phobia, agoraphobia, panic, overanxious, general anxiety, separation anxiety.

for the combined algorithm the lower of the two CGAS scores (either parent or youth) was applied in determining whether the CGAS ≤70 condition was met.

RESULTS

Comparison A: Test-Retest Agreement

Table 2 shows diagnostic agreement between the lay-administered DISC and the clinician-administered retest DISC. For each informant source, the three diagnostic algorithms described above have been applied.

Parent Informant. For the DISC-P, test-retest agreement for ADHD was good to excellent, with greatest agreement when the relatively more impaired cases were examined, using the most rigorous algorithm described above. For ODD, κ values were moderate, ranging from .65 to .73. There was moderate test-retest

agreement for CD (.56 to .59), with little difference by algorithm. For the depressive disorders, agreement was also moderate (.50 to .56) by all algorithms; for MDD, however, agreement was considerably better (.54 to .62) than for Dys (.30 to .38). For the anxiety disorders considered as a group, κ values ranged from .52 to .64, with a higher κ for the most impaired group. Of the three anxiety disorders examined individually, test-retest agreement was highest for OAD, whereas agreement for SAD and SoPh was marginal to poor.

Youth Informant. Test-retest agreement for ADHD and ODD was poor. For CD agreement was best (.64) for the criteria-only algorithm and lower when the algorithms requiring impairment were applied. For the depressive disorders, considered as a group and individually, test-retest agreement was marginal to poor. Agreement was uniformly poor for the anxiety disorders.

Combined Parent–Youth. For diagnoses drawing on combined parent and youth symptom reports, the fair to poor reliability of the child report pulled down the κ values somewhat (compared with κ values based on the parent report alone). An exception was CD, for which the criteria-only algorithm showed test-retest agreement of .66. The κ values for ADHD ranged from .48 to .64 and were .56 to .61 for ODD. Agreement was marginal for the depressive disorders (.37 to .45) and was fair for the anxiety disorders, with the highest levels of agreement for the most impaired cases.

Comparison B: DISC Validity with Retest

Table 3 shows diagnostic agreement between the lay-administered DISC and diagnoses generated by computer algorithm from symptom ratings made by the clinician after conducting the clinical-style interview.

Parent Informant. The κ values were generally good for ADHD and ODD, ranging from .57 to .82, with highest agreement in the subjects defined as more severely disturbed (CGAS ≤70). For CD, agreement was only fair, regardless of the algorithm used. For the depressive disorders, agreement was fair to good (.45 to .67). Agreement was fair for the anxiety disorders considered as a group (.44 to .53).

Youth Informant. Diagnostic agreement was poor except for CD, where κ ranged from .52 to .57 with only minor variability across algorithms.

Combined Parent–Youth. For diagnoses using symptom reports from both parent and youth, κ values are generally moderate for the externalizing disorders (.49 to .70). The κ values are in the mid .40s to lower .50s for the depressive disorders—slightly lower than for the parent report alone but considerably better than for the youth report alone. For the anxiety disorders as a group, levels of agreement are similar to those obtained from the parent report alone (.41 to .53).

Comparison C: DISC Concurrent Validity

Table 4 shows agreement between the clinician-administered DISC diagnoses and diagnoses generated from symptom ratings made after the clinical-style interview. Both sets of diagnoses were generated using computer algorithms.

Parent Informant. In the externalizing disorders, agreement was quite good for ADHD and CD (κ values greater than .70), while for ODD κ values were .56 to .59. There was little variability by diagnostic algorithm. Agreement was good for MDD (.55 to .60) but poor for Dys, resulting in fair levels of agreement (.45 to .49) for the composite depressive category. For the anxiety disorders, κ was .60 or greater for the composite category and slightly lower for OAD. The κ for SAD was poor; SoPh showed similarly poor agreement except for the criteria-only algorithm, which had a κ of .53.

Youth Informant. For ADHD, κ values were very low, indicating little more than chance agreement. For ODD, agreement was .54, regardless of algorithm. For CD, agreement was high for the criteria-only algorithm (.77) but only moderate for the algorithms that identified the more impaired youth. For the depressive disorders considered together and for MDD alone, agreement was high (.73 and above), while for Dys, κ values were .54 and .57. Agreement was only modest to poor for the anxiety disorders considered as a group (.31 to .49) and was best for SAD (.59). For OAD and SoPh considered separately, agreement was poor.

Combined Parent–Youth. For the externalizing disorders, κ values were similar to those for the parent report for ADHD (.70 to .72). The κ values were good (.62 to .65) for ODD, and they were good to high for CD (.65 to .80). For the depressive disorders as a group, κ values were in the moderate range (.51 to .57) and were substantially better for MDD than for Dys. For the anxiety disorders considered as a

TABLE 3

Comparison B: DISC Validity with Retest Lay DISC Diagnoses versus Diagnoses from Clinician Symptom Ratings

	Criteria Only		Criteria & Impairment		Criteria, Impairment & CGAS ≤70	
Diagnosis	κ	SE	κ	SE	κ	SE
Parent informant						
ADHD	.61	.063	.63	.063	.82	.064
ODD	.57	.063	.57	.063	.73	.064
CD	.48	.063	.48	.063	.49	.063
Depressive	.56	.062	.54	.062	.67	.064
MDD	.45	.064	.48	.063	.58	.062
Dys	.54	.062	.50	.062	.59	.063
Anxiety	.44	.063	.44	.064	.53	.062
OAD	.42	.063	.35	.064	.41	.064
SAD	.31	.060	.31	.060	.28	.044
SoPh	.33	.063	.32	.063	.38	.062
Youth informant						
ADHD	.24	.064	.18	.064	.22	.064
ODD	.33	.052	.26	.054	.30	.051
CD	.57	.063	.52	.064	.56	.064
Depressive	.33	.059	.22	.062	.29	.062
MDD	.39	.060	.27	.063	.35	.063
Dys	.33	.058	.32	.063	.32	.063
Anxiety	.33	.057	.33	.063	.43	.063
OAD	.27	.060	.35	.063	.48	.063
SAD	.32	.060	.32	.060	—	—
SoPh	.28	.061	.17	.062	.21	.063
Combined parent & youth						
ADHD	.49	.064	.51	.063	.62	.064
ODD	.56	.063	.54	.064	.61	.063
CD	.70	.063	.63	.063	.66	.063
Depressive	.53	.064	.48	.063	.51	.064
MDD	.48	.063	.45	.064	.49	.063
Dys	.50	.063	.52	.062	.53	.063
Anxiety	.41	.062	.46	.064	.53	.063
OAD	.35	.060	.44	.063	.58	.064
SAD	.32	.061	.32	.061	.36	.055
SoPh	.34	.064	.29	.059	.35	.062

Note: DISC = Diagnostic Interview Schedule for Children; ADHD = attention-deficit hyperactivity disorder; ODD = oppositional defiant disorder; CD = conduct disorder; MDD = major depressive disorder; Dys = dysthymia; OAD = overanxious disorder; SAD = separation anxiety disorder; SoPh = social phobia; Depressive = MDD or Dys; Anxiety = any of the following: social phobia, agoraphobia, panic, overanxious, general anxiety, separation anxiety.

group, agreement was moderate (.39 to .56); among the individual anxiety disorders, the best level of agreement was in OAD (.51).

DISCUSSION

This study has examined the criterion validity of the DISC, version 2.3, using a design that permitted different comparisons of DISC-generated diagnoses with diagnoses based on clinician symptom ratings. Generally the DISC has shown moderate to good validity across a number of diagnoses, although some areas of poor agreement with the clinical standard suggest that revision of questions for those diagnoses may be useful. In particular, the anxiety disorders and dysthymia warrant consideration for revision as they showed relatively lower reliability and validity. In considering the findings reported here, three main sources

TABLE 4
Comparison C: DISC Concurrent Validity of Clinician-Administered DISC Diagnoses versus Diagnoses from Clinician Symptom Ratings

	Criteria Only		Criteria & Impairment		Criteria, Impairment & CGAS ≤70	
Diagnosis	κ	SE	κ	SE	κ	SE
Parent informant						
ADHD	.72	.063	.71	.063	.72	.063
ODD	.59	.063	.56	.063	.57	.063
CD	.74	.063	.74	.063	.74	.063
Depressive	.48	.059	.45	.058	.49	.058
MDD	.60	.063	.55	.062	.60	.062
Dys	.35	.059	.37	.058	.38	.058
Anxiety	.62	.063	.60	.064	.64	.064
OAD	.60	.063	.57	.062	.56	.063
SAD	.29	.059	.33	.061	.38	.060
SoPh	.53	.062	.38	.063	.39	.063
Youth informant						
ADHD	.27	.058	.18	.036	.19	.038
ODD	.54	.061	.54	.061	.54	.061
CD	.77	.064	.45	.060	.52	.060
Depressive	.77	.064	.73	.061	.73	.061
MDD	.79	.063	.73	.062	.73	.062
Dys	.54	.063	.57	.057	.57	.057
Anxiety	.49	.064	.31	.056	.41	.056
OAD	.23	.062	.32	.047	.36	.049
SAD	.59	.062	—	—	—	—
SoPh	.45	.063	.20	.049	.25	.053
Combined parent & youth						
ADHD	.70	.063	.72	.062	.72	.062
ODD	.65	.063	.62	.062	.62	.062
CD	.80	.063	.65	.062	.65	.062
Depressive	.57	.061	.51	.059	.54	.059
MDD	.63	.062	.57	.061	.60	.061
Dys	.37	.061	.37	.058	.37	.058
Anxiety	.56	.064	.39	.061	.50	.062
OAD	.51	.064	.46	.061	.49	.062
SAD	.40	.061	.39	.063	.48	.061
SoPh	.43	.063	.29	.055	.33	.057

Note: DISC = Diagnostic Interview Schedule for Children; ADHD = attention-deficit hyperactivity disorder; ODD = oppositional defiant disorder; CD = conduct disorder; MDD = major depressive disorder; Dys = dysthymia; OAD = overanxious disorder; SAD = separation anxiety disorder; SoPh = social phobia; Depressive = MDD or Dys; Anxiety = any of the following: social phobia, agoraphobia, panic, overanxious, general anxiety, separation anxiety.

of variability in DISC–clinician diagnostic agreement are evident, in addition to that attributable to the instrument itself.

First, findings on the validity of the DISC vary depending on the informant used (parent, youth, both). With a few exceptions (e.g., depressive disorders in comparison C), diagnoses based on the youth report appear less valid than those based on the parent report. This pattern of poorer validity with youth informants is particularly evident in comparison B for most diagnoses (except CD) and in comparison C for ADHD, CD with impairment, and the anxiety disorders (except for SAD). For the youth report of anxiety disorders, this general pattern of marginal to poor agreement between diagnoses from the DISC and from a clinically based assessment is in accord with findings from other studies

that show generally poor concordance for child-based reports of anxiety disorders "across all interviews and conditions" (Hodges, 1993). The need for an adult informant to achieve a reasonably valid assessment of ADHD has also been noted (Hodges, 1993), and it is likely that the same applies to ODD. These two disruptive behavior disorders require reporting on the child's negative impact on the social environment, a perspective that may be difficult for a youth to report accurately. By contrast, the symptoms of CD have less to do with the interpretation of interpersonal impact; they are more likely to be transgressions of social norms that are often matters of fact (e.g., events) and sometimes of public record. Reporting on them is a less subjective matter, and hence one on which there may be greater agreement across assessments. Another potential reason for the generally lower levels of concordance for youth reports involves potential developmental effects in symptom reporting. Children 11 years and younger have been shown to be fairly unreliable in their symptom reporting (Edelbrock et al., 1985; Fallon and Schwab-Stone, 1994; Schwab-Stone et al., 1994), and about one third of the sample in this study was in the 9- through 11-year age range.

A second source of variability in the findings reported here lies in the design of the comparison between DISC and clinician-derived diagnoses. In anticipation of this effect, the study was designed to allow examination of test-retest agreement for the DISC alone, since limitations in agreement at this level necessarily reduce concordance between the diagnoses from the first DISC assessment and those based on clinical assessment occurring later (i.e., at retest). It is evident that there is some limitation on diagnostic concordance after an interval (Table 2), particularly for the youth informant. When agreement between DISC and clinician-derived diagnoses is examined for the component of the study in which the retest interval was in effect eliminated (comparison C), concordance for the youth report is substantially better across most comparisons (except for ADHD, OAD, and SoPh), and this improvement is more striking among the youth than among the parent informants. With respect to the study design, it can be argued that comparison C does not provide two truly independent diagnostic evaluations, one of which is the "standard." On the other hand, it can also be argued that the structure inherent in the administration of the DISC constrains the clinician, making that assessment reasonably comparable with the lay DISC interview and relatively unlike that which occurs when the clinician is allowed to question freely and apply clinical judgment in determining whether a symptom is present or not. Following this line of argument, comparison C, by reducing retest effects, allows a more precise evaluation of whether the DISC interview process elicits symptom-related information and assigns it to the same symptom status as clinicians do. Nevertheless, it is clear from the test-retest κ values reported in Table 2 and from comparing results from comparison B with those from comparison C (Table 3 versus Table 4) that reporting effects over time (usually attenuation in symptom reporting) (see Jensen et al., 1995) are operating to reduce DISC–clinician diagnostic concordance, particularly for the youth report.

A third source of variability in the levels of diagnostic agreement derives from the algorithms applied to synthesize symptom reports into *DSM-III-R* diagnoses. It was expected that by adding requirements of diagnosis-specific impairment and then also of CGAS $\leq$70 to the rules for determining the presence of a diagnosis, a relatively more severe group of cases would be identified, and concordance with clinical assessment would improve over that found when the "criteria only" algorithms were applied. The application of CGAS $\leq$70 to the algorithm led to some improvement in test-retest agreement and validity as assessed by comparison B for the parent report of ADHD and ODD, as well as some improvement in comparison B validity for the parent report of depressive disorders; however, overall there was not a great impact from the inclusion of requirements for impairment in the diagnostic algorithms. Although the reasons for this lack of effect are not clear at this point, a number of potential explanations deserve further exploration. These include the possibility that unreliability in the impairment questions was sufficient to offset potential gain from delimiting a group of subjects with more severe disturbances. Another possibility is that the DISC revision process (Shaffer et al., 1993) has led to question wordings that set a relatively higher threshold for positive response (i.e., that are no longer "oversensitive") (see Shaffer et al., 1993, 1996). If this were the case, requiring specific impairment criteria in addition to the diagnostic criteria might not provide sufficient restriction to make a noticeable difference in levels of agreement.

In summary, diagnostic validity for the DISC-2.3 ranges from moderate to very good when parent-reported information is considered alone and also when parent and youth reports are combined. Validity of diagnoses based on the youth report alone is particularly subject to retest effects and is generally fair to poor in the comparison requiring retest; however, when retest is not required, concordance between DISC and clinician-derived diagnoses is moderate for some comparisons and is high for MDD.

In arriving at an overall appraisal of the validity of the DISC at this stage in its development, it would be useful to compare it with the other diagnostic interviews used in child psychiatric research. The only comparable study is that conducted by Boyle et al. (1993) which examined the validity of the Diagnostic Interview for Children and Adolescents-Revised (DICA-R) in a community sample of 32 youths (oversampled for symptomatology) using a design similar to the one reported here. Using a retest design, lay-administered DICA-R diagnoses were compared with diagnoses derived from clinician ratings after clinical-style probing. The κ values were generally very good (i.e., most in the range .55 to .84). In comparing these findings with results from the current study, two basic differences in the instruments must be considered. First, the DICA-R that was used by Boyle et al. (1993) assessed only present diagnoses rather than diagnoses from the previous 6 months, which is the time frame for DISC-2.3 symptom reporting. Second, as noted by Shaffer et al. (1994), the DISC-2.3 has been constructed to adhere very tightly to *DSM-III-R* criteria, including those that cause the most difficulty for respondents, i.e., frequency, duration, and co-occurrence. The DICA-R does not follow quite the same structure, but relies more heavily on symptom criteria; thus, in certain instances the DICA-R functions somewhat like a diagnosis-specific symptom scale (Shaffer et al., 1994). The effect of this structure is to increase prevalence (Boyle et al., 1993) and to enhance concordance between lay and clinician assessments (i.e., see reliability of age-of-onset items in Shaffer et al., 1996). Thus, even fairly subtle differences in study design, algorithms, time frame, and sample must be considered in evaluating findings relevant to the validity of structured interviews. Other recent validity studies have utilized clinic referrals, outpatients, and inpatients (see Piacentini et al., 1993), with the exception of the Cohen et al. (1987) study which used an early version of the DISC and found only modest agreement with the clinical standard. A major strength of the current study lies in its use of a substantial community sample.

Another potential comparison for these results is with the Diagnostic Interview Schedule (DIS) (Robins et al., 1981), which was developed and evaluated in the context of the adult Epidemiologic Catchment Area studies. Results from a study conducted with a large sample of adult community respondents at the St. Louis site using a lay-clinician, test-retest design yielded κ values for lifetime diagnostic agreement which ranged from 0.12 to 0.63 (Helzer et al., 1985), while the Baltimore-based clinical reappraisal study (Anthony et al., 1985) found very low to modest levels of agreement between DIS and clinician diagnoses (e.g., κ values = $-.02$ to .35). It is encouraging that the DISC-2.3 compares favorably with the major structured diagnostic interview for epidemiological use with adults, despite the difficulties of evaluating psychopathology across a broad developmental range and from the perspectives of two informants. It is certainly fair to say that steady progress has been made in the development of valid strategies for the psychiatric assessment of children and adolescents in community studies. Nevertheless, these results also raise questions—about the nature of youth symptom reports, their most appropriate use, potential developmental effects, and the integration of impairment and symptom reports—which remain as challenges for future research in child diagnostic assessment.

The MECA Program is an epidemiolgical methodology study performed by four independent research teams in collaboration with staff of the Division of Clinical Research, which was reorganized in 1992 with components now in the Division of Epidemiology and Services Research and the Division of Clinical and Treatment Research, of the NIMH, Rockville, MD. The NIMH Principal Collaborators are Darrel A. Regier, M.D., M.P.H., Ben Z. Locke, M.S.P.H., Peter S. Jensen, M.D., William E. Narrow, M.D., M.P.H., Donald S. Rae, M.A., John E. Richters, Ph.D., Karen H. Bourdon, M.A., and Margaret T. Roper, M.S. The NIMH Project Officer was William J. Huber. The Principal Investigators and Coinvestigators from the four sites are as follows: Emory University, Atlanta, U01 MH46725: Mina K. Dulcan, M.D., Benjamin B. Lahey, Ph.D., Donna J. Brogan, Ph.D., Sherryl H. Goodman, Ph.D., and Elaine W. Flagg, Ph.D.; Research Foundation for Mental Hygiene at New York State Psychiatric Institute, New York, U01 MH46718: Hector R. Bird, M.D., David Shaffer, M.D., Myrna Weissman, Ph.D., Patricia Cohen, Ph.D., Denise Kandel, Ph.D., Christina Hoven, Ph.D., Mark Davies, M.P.H., Madelyn S. Gould, Ph.D., and Agnes Whitaker, M.D.; Yale University, New Haven, CT, U01 MH46717: Mary Schwab-Stone,

M.D., Philip J. Leaf, Ph.D., Sarah Horwitz, Ph.D., and Judith H. Lichtman, M.P.H.; University of Puerto Rico, San Juan, U01 MH46732: Glorisa Canino, Ph.D., Maritza Rubio-Stipec, M.A., Milagros Bravo, Ph.D., Margarita Alegria, Ph.D., Julio Ribera, Ph.D., Sara Huertas, M.D., Michael Woodbury, M.D., and Jose Bauermeister, Ph.D.

REFERENCES

Anthony J, Folstein M, Romanoski A et al. (1985), Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis. *Arch Gen Psychiatry* 42:667–675

Boyle MH, Offord DR, Racine Y et al. (1993), Evaluation of the Diagnostic Interview for Children and Adolescents for use in general population samples. *J Abnorm Child Psychol* 21:663–681

Bravo B, Woodbury-Farina M, Canino GJ, Rubio-Stipec M (1993), The Spanish translation and cultural adaptation of the Diagnostic Interview Schedule for Children (DISC) in Puerto Rico. *Cult Med Psychiatry* 17:329–344

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

Cohen P, O'Connor P, Lewis S, Velez N, Malachowski B (1987), Comparison of DISC and K-SADS-P interviews of an epidemiological sample of children. *J Am Acad Child Adolesc Psychiatry* 26:662–667

Costello AJ, Edelbrock CS, Dulcan MK, Kalas R, Klaric SH (1984), Development and Testing of the NIMH Diagnostic Interview Schedule for Children (DISC) in a Clinic Population. Final Report (contract RFP-DBB-81–0027). Rockville, MD: Center for Epidemiological Studies, NIMH

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

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

Fisher P, Shaffer D, Piacentini J et al. (1993), Sensitivity of the Diagnostic Interview Schedule for Children, 2nd edition (DISC 2.1) for specific diagnoses of children and adolescents. *J Am Acad Child Adolesc Psychiatry* 32:666–673

Fleiss JL (1981), *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley

Helzer J, Robins L, McEvoy L et al. (1985), A comparison of clinical and diagnostic interview schedule diagnoses. *Arch Gen Psychiatry* 42:657–666

Hodges K (1993), Structured interviews for assessing children. *J Child Psychol Psychiatry* 34:49–68

Jensen P, Roper M, Fisher P et al. (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

Lahey BB, Flagg EW, Bird HR et al. (1996), The NIMH Methods for the Epidemiology of Child and Adolescent Mental Disorders (MECA) Study: background and methodology. *J Am Acad Child Adolesc Psychiatry* 35:855–864

Piacentini J, Shaffer D, Fisher P, Schwab-Stone M, Davies M, Gioia P (1993), The Diagnostic Interview Schedule for Children-Revised Version (DISC-R): III. Concurrent criterion validity. *J Am Acad Child Adolesc Psychiatry* 32:658–665

Ribera J, Canino G, Bravo M et al. (1996), Diagnostic Interview Schedule for Children (DISC-2.1) in Spanish: reliability in a Hispanic population. *J Child Psychol Psychiatry* 37:195–204

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

Robins L, Helzer J, Croughan J, Ratcliff K (1981), National Institute of Mental Health Diagnostic Interview Schedule. *Arch Gen Psychiatry* 38:381–389

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, Bird H, Bourdon K et al. (1994), DISC 2.3: a summary of its performance in the MECA study, comparisons with other diagnostic instruments, and an update on the DISC 3.0. Paper presented at the Annual Meeting of the Society for Research in Child and Adolescent Psychopathology, June

Shaffer D, Fisher P, Dulcan MK et al. (1996), The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): description, acceptability, prevalence rates, and performance in the MECA study. *J Am Acad Child Adolesc Psychiatry* 35:865–877

Shaffer D, Gould MS, Brasic J et al. (1983), A children's global assessment scale (CGAS). *Arch Gen Psychiatry* 40:1228–1231

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