Test-Retest Reliability of the Diagnostic Interview Schedule for Children (DISC 2.1)

Peter Jensen, National Institute of Mental Health
Margaret Roper, National Institute of Mental Health
Prudence Fisher, Columbia University College of Physicians and Surgeons
John Piacentini, Columbia University College of Physicians and Surgeons
Glorisa Canino, University of Puerto Rico
John Richters, National Institute of Mental Health
Maritza Rubio-Stipec, University of Puerto Rico
Mina Dulcan, Emory University
Sherryl H Goodman, Emory University
Mark Davies, Columbia University College of Physicians and Surgeons

Only first 10 authors above; see publication for full author list.

Journal Title: Archives of General Psychiatry
Volume: Volume 52, Number 1
Publisher: American Medical Association (AMA) | 1995-01-01, Pages 61-61
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1001/archpsyc.1995.03950130061007
Permanent URL: https://pid.emory.edu/ark:/25593/rfm56

Final published version:
http://dx.doi.org/10.1001/archpsyc.1995.03950130061007

Copyright information:
© 1995, American Medical Association

Accessed March 30, 2019 2:33 PM EDT
Test-Retest Reliability of the Diagnostic Interview Schedule for Children (DISC 2.1)

Parent, Child, and Combined Algorithms

Peter Jensen, MD; Margaret Roper, MS; Prudence Fisher, MS; John Piacentini, PhD; Glorisa Canino, PhD; John Richters, PhD; Maritza Rubio-Stipec, MA; Mina Dulcan, MD; Sherryl Goodman, PhD; Mark Davies, MPH; Donald Rae, MS; David Shaffer, MD; Hector Bird, MD; Benjamin Lahey, PhD; Mary Schwab-Stone, MD

Background: Previous research has not compared the psychometric properties of diagnostic interviews of community samples and clinically referred subjects within a single study. As part of a multisite cooperative agreement study funded by the National Institute of Mental Health, 97 families with clinically referred children and 278 families identified through community sampling procedures participated in a test-retest study of version 2.1 of the Diagnostic Interview Schedule for Children (DISC 2.1).

Methods: The DISC was separately administered to children and parents, and diagnoses were derived from computer algorithms keyed to DSM-III-R criteria. Three sets of diagnoses were obtained, based on parent information only (DISC-P), child information only (DISC-C), and information from either or both (DISC-PC).

Results: Test-retest reliability of the DISC-PC ranged from moderate to substantial for diagnoses in the clinical sample. Test-retest k coefficients were higher for the clinical sample than for the community sample. The DISC-PC algorithm generally had higher reliabilities than the algorithms that relied on single informants. Unreliability was primarily due to diagnostic attenuation at time 2. Attenuation was greatest among child informants and less severe cases and in the community sample.

Conclusions: Test-retest reliability findings were consistent with or superior to those reported in previous studies. Results support the usefulness of the DISC in further clinical and epidemiologic research; however, closely spaced or repeated DISC interviews may result in significant diagnostic attenuation on retest. Further studies of the test-retest attenuation phenomena are needed, including careful examination of the child, family, and illness characteristics of diagnostic stability.

(Arch Gen Psychiatry. 1995;52:61-71)
METHODS

The data reported were obtained during the first phase of the MECA Study. The DISC 2.1 was administered as part of an extensive battery of measures being tested, including demographic characteristics, global impairment, service use, barriers to service use, potential presumed risk factors, and correlates of childhood psychopathologic conditions.

SAMPLE

The subjects were 97 clinically referred and 278 community-sampled 9- to 17-year-old children and one of their primary caretakers (usually the mother). Clinical subjects were recruited from clinical settings in three geographic areas (Atlanta, Ga; New York, NY; and Puerto Rico) involved in the MECA Study. The 278 nonclinical subjects were recruited across the three sites from community populations. For these community subjects, one site (New York) used a stratified random sampling procedure (with replacement of refusals), and two sites (Atlanta and Puerto Rico) used household probability sampling procedures (with no replacement). Data from a fourth site (Connecticut) are not presented here, since this site used different procedures and examined validity rather than reliability. The three sites' sampling procedures are described below.

New York

The New York sample was drawn from an area directly north of New York City, where there is a population with significant socioeconomic, geographic, and ethnic diversity, including sizable proportions of blacks, Hispanics, and other ethnic minorities. The aims of the New York site's sampling procedure were to achieve heterogeneity and to parallel the general population's proportions in age, sex, socioeconomic status (SES), population density, and ethnicity.

Fifty-three of the 63 subjects from the clinical sample were outpatients from school-based outpatient clinics in five counties within the designated catchment area. Ten additional participating subjects (from the same region) were currently hospitalized at a child psychiatric facility. All of the clinical subjects were active cases and had been involved with the clinical facility for a minimum of 3 months and for a maximum of 2 years. The clinical subjects were drawn from a larger sample of 123 eligible clinical subjects, resulting in a study participation rate of 51%. The community sample was obtained by using Department of Education data tapes to determine the age and ethnicity profiles of school systems, within which individual schools were selected based on their representativeness of the characteristics of the community. Within target schools, English classes (or “homeroom” classes for younger ages) were selected at random, as were students within these classes, based on the stratification target numbers. Children enrolled in classes for English as a second language and those in special education classes were excluded so that responses would not be biased by lack of language proficiency or cognitive ability. Three hundred sixty-four eligible subjects were identified, of whom the parents of 138 children indicated by letter or telephone that they did not wish to participate (38%). An additional 104 families could not be reached to secure study participation, and complete data were not available for two families. Despite these difficulties, the resulting sample (n=120; study participation rate, 33%) achieved the site's sampling aims for heterogeneity, applied to the socioeconomic spectrum and the age groups studied, consisted of half white and half nonwhite subjects, and paralleled the general population for urban/suburban/rural status. There was a preponderance of females in the community sample, in contrast to more males in the clinical sample (Table 1).

Puerto Rico

The clinical test-retest subjects were obtained from 9- to 17-year-olds consecutively admitted to an outpatient clinic (18 cases) and additional subjects (10 cases) from a substance abuse treatment center. Two thirds of subjects were male. Twenty-four of the 28 cases completed the interviews at both time points. As expected in clinical samples drawn from the public sector, children in the clinic sample were predominantly male (Table 1) and of low SES (82%). Refusal rates were not tracked, so it was not possible to examine factors that may have shaped study participation in this sample.

Community subjects were obtained from a probability sample of children ages 9 through 17 years from the San Juan metropolitan area. Sampling blocks were randomly distributed into two community subsamples of 124 subjects each to carry out two separate studies designed to measure reliability and validity of the DISC. This report concerns only one of the two samples (the reliability sample). Of these 124 families, 118 pairs (95%) agreed to participate and completed the study. This community sample resembled the 1990 census population of children in Puerto Rico ages 9 to 17 years, evenly distributed by gender, with most of low SES.

Atlanta

Twelve children being seen in a psychiatric outpatient clinic were selected for participation in the test-retest study; of these 12, one family refused (92% participation). As in the New York and Puerto Rico clinical samples, males outnumbered females. The community sample was selected from two census tracts in Georgia, one in an urban county and one in a rural county. These two census tracts were chosen so that the resultant probability sample of youths would have variation in race, urban/rural residence, and SES. A stratified multistage survey design was used to identify youths to be interviewed. Sampling regions in the urban census tract were stratified by race (mostly black, mostly white, mixed) and by SES (lower, upper). The rural census tract was stratified by race only since the SES appeared relatively homogeneous. Differential sampling rates were applied across strata to equalize the number of black and white youths and the number of lower and upper SES youths in the sample. Sample housing units were randomly selected for participation in the survey. If there was more than one child of the appropriate age living in the household, one was randomly selected using a table of random numbers. Sixty-six subjects were selected for study participation; of these, 26 refused, leaving 40 subjects and families (61%) who participated in the community test-retest study at the Atlanta site.

INSTRUMENT

The DISC 2.1, a highly structured diagnostic instrument intended for lay administration, was administered to parents and
...children. Parent and child versions of the interview require approximately 60 to 75 minutes of the subject's time. This version of the DISC (version 2.1) differs significantly from earlier versions of the instrument. In that items were grouped into separate diagnostic modules, each based on a set of related diagnoses. Also, the time covered by the interview was changed from 1 year to 6 months (to maintain consistency with the duration requirements in DSM-III-R), and questions concerning age at first episode, current impairment, treatment history, and precipitating stressors were added at the end of each diagnostic module. Further specific refinements were made based on the input of experienced clinical investigators concerning exact wording of specific questions. Also, unreliable items were deleted, based on data analyses from the earlier DISC-Revised field trials. 

All DISCs were administered by college or graduate level "lay" interviewers who had completed an intensive training period (usually 2 weeks, with slight site-to-site variation). The DISC interviews were monitored by experienced child interviewers and/or child clinicians through review of audi-tapes. After the interview was completed, all responses from both parents and children were entered into a computer. Using computer algorithms based on DSM-III-R diagnostic criteria, DSM-III-R diagnoses were generated separately from parent information (DISC-P) and child information (DISC-C). The algorithms also generated a third set of diagnoses based on the "or" rule (DISC-PC, where PC indicates both parent and child informants). The "or" rule specifies that either informant may provide symptom information, which when combined may be used to jointly meet criteria for a particular diagnosis, even though the diagnostic criteria may not be met by any single informant. Thus, the "or" rule accepts the endorsement of a positive symptom from either the child or the adult informant, while the determination of symptom absence requires two negative responses. The application of the "or" rule generally tends to inflate prevalence but is thought to more closely approximate clinical practice.

PROCEDURES

Clinically referred cases were generally interviewed in the clinical setting, while community subjects were interviewed in their homes. In general, retest intervals were shorter for clinical than community subjects (about 2 weeks vs 3 weeks), and sites differed significantly in test-retest intervals (Table 1). All interviews (parent, child, and test-retest) were conducted by separate interviewers who had no knowledge of the other interviews, requiring a total of four different interviewers for each parent-child pair. The use of two different interviewers for each administration introduces maximum variability of conditions and submits the instrument to a stringent test of its reliability. Parent and child interviews were conducted concurrently (at the same time but in separate rooms) for over 93% of cases. The adult who knew the selected child best (usually the mother) was chosen as the adult informant.

DATA ANALYSIS

There currently exists little empiric information that demonstrates superior validity for either parent or child information. The current standard of clinical diagnostic approaches usually involves gathering information from both the parent and the child. Accordingly, we restricted our initial analyses to the combined (DISC-PC) algorithm results. Research support for this approach is provided in several recent studies that have indicated that statistical weighting of informants' responses or exclusive reliance on single-informant algorithms does not perform as well as the simple combinatorial "or" rule against clinicians' diagnoses.

The data analytic strategy proceeded as follows: First, reliability statistics from individual sites' data sets were compared to determine whether sites differed in reliability. This was warranted as an initial strategy given the geographic, linguistic, and sampling differences between sites. Furthermore, given the instability of k statistics with few cases, site differences were examined only for those diagnoses for which five or more subjects met diagnostic criteria at the initial time point at two or more sites. In some instances, all three sites had sufficient cases to allow this comparison; in other instances, only two of the three sites had five or more cases; and in several instances, sites could not be compared because of insufficient numbers.

Because we found few or no site differences, further analyses of combined sites' data were undertaken. In most instances, clinical and community samples were analyzed separately. Except where noted, diagnostic reliabilities were examined for DISC-PC diagnoses. Given sufficient cases, diagnostic reliabilities were compared between males and females, younger and older children, internalizing vs externalizing disorders, clinic vs community subjects, and parent vs child informants. Some diagnoses may be more stable as a function of sex, considering the sex differences in prevalence of most childhood conditions, both before and after puberty. Similarly, some diagnoses may be more stable as a function of age, considering the age differences in prevalence of most childhood conditions.

Given the relatively small sample size and number of positive diagnoses, priority was given to examining test-retest reliability for the most common childhood conditions. There were sufficient cases of ADHD, ODD, and CD to allow for the examination of their reliabilities as separate categories. For most affective and anxiety disorders, however, too few cases were available to conduct meaningful reliability analyses across the clinical and community samples or to conduct more fine-grained analyses by age and sex. Therefore, we constructed two additional diagnostic categories: (1) Children with major depression and/or dysthymia were combined into a single depressed diagnostic group. (2) Children with any of the DSM-III-R anxiety disorder diagnoses (simple phobia, social phobia, separation anxiety disorder, generalized anxiety disorder, panic disorder, overanxious disorder, etc) were combined into a single group with any anxiety disorder. More detailed analyses of specific affective and anxiety disorders (e.g., major depression, simple phobia, and separation anxiety disorder) are available from the authors on request.

To parallel the major conceptual distinction drawn by many investigators concerning types of childhood psychopathologic conditions, we also constructed superordinate categories of any internalizing disorder (major depressive disorder, dysthymia, and/or any anxiety disorder) and any externalizing disorder (ADHD, ODD, and CD). Reliability coefficients were computed for each of the three diagnostic algorithms (DISC-P, DISC-C, and DISC-PC).

To examine the possibility that unreliability is related to the relative oversensitivity of k statistics to slight differences

Continued on next page
in symptom levels around the diagnostic threshold, we computed intraclass correlations based on the test-retest diagnostic criteria for all major diagnostic categories. For example, the 14 ADHD symptoms listed in the diagnostic criteria were treated as a dimensional scalar variable, and the intraclass correlation was computed between the two time points for this “scale.” The criteria for most diagnoses involve some degree of summated symptoms as well as duration and age of onset items. However, because onset and duration criteria rarely can be summed for more than two items, intraclass correlations were only computed for symptom count criteria.

tured lay diagnostic interviews, multi-informant approaches, risk factor assessments, etc) in a moderated-sized survey of children and adolescents prior to mounting a full-scale nationwide study.

Further development of the DISC is forthcoming, pending analyses from the two field trials conducted during the MECA Study. The validity of the major diagnostic categories is being assessed as a part of the second MECA field trial and will form the basis of a future report. The focus of this report concerns the first MECA field trial, during which a test-retest reliability study was carried out. Information concerning the test-retest characteristics of the instrument is critical to assist field workers and clinical researchers who wish to conduct studies with this instrument.

Previous studies examining the reliability of earlier versions of the DISC reported moderate test-retest reliabilities, with the best reliabilities in the externalizing disorders (conduct disorder [CD], attention-deficit hyperactivity disorder [ADHD], oppositional defiant disorder [ODD]). These studies were carried out only in clinical settings, an important limitation, since instruments’ psychometric properties are best examined in settings similar to those where they will be used. Instruments behave quite differently in clinical vs community settings; they will likely be less reliable and less valid in the community. Testing the psychometric performance of the agreed-on diagnostic interview in community settings was a central stated goal of the initial request for applications for the MECA Study. The DISC was the diagnostic instrument chosen by the MECA investigators.

Unfortunately, few systematic studies exist concerning the reliability of any child diagnostic instrument in community settings, so points of comparison are difficult to establish. Indeed, we have been unable to locate any studies that have compared the psychometric properties of diagnostic interviews for community samples and clinically referred subjects within a single study, either of children or of adults. Some relevant information bearing on this question is found in the adult literature, however. As a part of the Epidemiologic Catchment Area Study, Helzer et al. examined lay-clinician test-retest reliability in a community sample. They found lower reliabilities with the Diagnostic Interview Schedule (DIS) for adults than reported for clinical samples assessed with earlier versions of the instrument. Similar findings have been noted by other investigators, with low reliability, diagnostic stability, and validity in community subjects compared with clinical subjects.

If diagnostic instruments are less reliable in community settings than in clinical settings, the discrepancy may be due to several factors. First, those with severe cases may be more likely to seek out or be referred for care than those with mild cases, who may hover just above the diagnostic “threshold.” Among such threshold subjects, small decreases in symptom reporting over the test-retest interval will result in decreased reliability. Second, as clinical subjects and parents learn more about a diagnosis and/or its symptoms during the evaluation or treatment process, they may become more accepting of the symptoms and diagnostic process and more likely to reaffirm the presence of symptoms in a test-retest paradigm. Third, parents of children referred to clinical settings may have justifiably greater concern about their children and may respond more carefully to interview questions than parents from a community sample, whose motivation to participate is less well understood. Also, given two informants, the parent and child could confer about their interview responses between the first test and the retest. It is unclear what effects (if any) the possibility of parent-child discussions may have on retest, but some instances of attenuation may be explained by this factor.

The purpose of this report from the MECA Study is to describe the test-retest reliability of the recently revised DISC (version 2.1) in clinical settings and to extend previous research by also examining its test-retest reliability in community settings across the participating MECA sites. Factors mediating differences between clinical and community sample test-retest K statistics are explored.

**RESULTS**

In Table 1 we outline the number of subjects from the clinical and community settings from each site as well

---

**Table 1. Site Characteristics**

<table>
<thead>
<tr>
<th>Site</th>
<th>No. of Pairs in Test-Retest</th>
<th>Average (SD) Age, y</th>
<th>M/F Ratio</th>
<th>Mean (SD) Retest Interval, d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinic Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td>62</td>
<td>12.1 (2.2)</td>
<td>47/15</td>
<td>14.7 (8.4)</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>24</td>
<td>14.3 (2.3)</td>
<td>16/6</td>
<td>10.0 (7.5)</td>
</tr>
<tr>
<td>Georgia</td>
<td>11</td>
<td>12.2 (2.8)</td>
<td>10/1</td>
<td>5.8 (6.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F = 5.8</td>
<td>df = 1.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>df = 2</td>
<td>P = .001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community Subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td>120</td>
<td>13.4 (2.6)</td>
<td>51/69</td>
<td>19.0 (13.5)</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>118</td>
<td>12.6 (2.4)</td>
<td>59/59</td>
<td>21.3 (13.0)</td>
</tr>
<tr>
<td>Georgia</td>
<td>40</td>
<td>13.2 (2.9)</td>
<td>19/21</td>
<td>33.3 (32.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F = 3.02</td>
<td>df = 1.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>df = 2</td>
<td>P = .06</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
as the mean age of subjects, male-female ratio, and test-retest intervals. Sites differed significantly in terms of age, with Puerto Rican subjects significantly older than subjects from the other two sites in the clinical setting. In contrast, Puerto Rican subjects were significantly younger than subjects from the other two sites in the community setting. Sites did not differ in proportions of males and females. There was a preponderance of males in the clinical setting, possibly reflecting the increased referral rates of young males commonly found in studies of clinical populations.20

Given the range of methods across sites and the differences in age noted above, in Table 2 we considered the extent to which findings differed across sites. Given the relatively modest power to test for site differences because of the small number of cases, we did not correct for the number of comparisons in this table. Despite the methodologic, geographic, cultural, linguistic, and ethnic differences among sites and despite the fact that the DISC had been translated into Spanish for use in Puerto Rico, there was little evidence of consistent or large intersite differences in either the clinical or community setting. Therefore, data for individual sites were combined for all subsequent analyses.

Table 3 provides information on the test-retest reliability of the DISC-PC in the clinical and community settings. Extensive information is not provided on the parent and child algorithms because the DISC-PC generally yielded more stable and higher reliabilities than the DISC-P or DISC-C. Comparisons of clinical and community values indicate that, as predicted, the χ coefficients are higher in clinical than community subjects (five of five χ coefficients were higher in the clinical sample, P=.05 by one-tailed sign test). This same pattern held within each site. Further comparisons indicated that at the individual diagnostic level, clinical χ coefficients were higher than community χ coefficients for depression and/or dysthymia (.70 vs .26, χ²=15.3, P=.0001). A similar but nonsignificant trend was noted for any anxiety disorder (.50 vs .32, χ²=2.45, P=.11). Kappa coefficients for the other diagnoses did not differ between clinic and community settings (analyses available from the authors on request).

Table 4 provides further information about the test-retest reliability of the DISC-P and DISC-C. Among community subjects, four of five χ coefficients were higher for DISC-P than for DISC-C, and among clinical subjects, four of five χ coefficients were higher for DISC-P than for DISC-C.

Analyses were conducted to determine whether diagnostic reliabilities were related to the child’s gender (10 comparisons, five for each major diagnostic area, with separate analyses for community and clinic subjects). These analyses yielded only one significant difference in reliability as a function of gender: within the community sample, higher test-retest reliability was found for depression and/or dysthymia among female subjects (k=.40, N=148) than among male subjects (k=.01, N=126) (χ²=23.5, P=.0001).

To determine whether reliabilities were related to the child’s age, the samples were divided into subjects 9 through 12 years and those 13 through 18 years. This split allowed the most even distribution of cases in the two age groups and is reasonably consistent with age differences in diagnoses before and after puberty. These analyses (10 comparisons, five for each major diagnostic area, with separate analyses for community and clinic subjects) yielded two significant differences as a function of age: Within the community sample, higher test-retest reliability was found for depression and/or dysthymia among older subjects (k=.28, N=152) than among younger subjects (k=.00, N=117). Also within the community sample, higher test-retest reliability was found for conduct disorder among older subjects (k=.68, N=150) than among younger subjects (k=.43, N=118) (χ²=5.19, P=.02). Given the small number of cases and relatively modest power to find true differences, the number of comparisons was not corrected for chance and should be viewed with caution.

To assess the possibility that unreliability is related to the sensitivity of χ statistics to slight differences in symptom or criterion levels around the diagnostic thresh-
old, intraclass correlations were computed to examine the test-retest reliability for the diagnostic criteria for all major diagnostic categories. Intraclass correlation coefficients for clinical subjects were .63 for any anxiety disorder, .68 for depression and/or dysthymia, .79 for ADHD, .76 for ODD, and .88 for CD. Similarly, community test-retest intraclass correlation coefficients were .38 for depression and/or dysthymia, .47 for any anxiety disorder, .74 for ADHD, .63 for ODD, and .68 for CD. As can be seen when these figures are compared with those in Table 3, in all instances but one the intraclass correlation coefficients were higher than the $\kappa$ coefficients, and these differences were sizable, especially for community subjects. These data are supportive of the hypothesis that some difficulties with reliability occur around the diagnostic threshold, particularly in community cases. Given that the intraclass correlations themselves fell significantly short of optimal reliability estimates, however, it must be concluded that threshold cases at time 1 are not a sole or sufficient explanation for test-retest attenuation or the less-than-optimal reliabilities.

**SOURCES OF UNRELIABILITY**

Given the lower reliabilities found in community settings (especially for anxiety and depressive disorders), we conducted a number of post hoc analyses to further explore the sources of unreliability and attenuation. We examined the extent to which caseness unreliability might be related to symptom severity at time 1, so that cases less severely impaired at time 1 were less likely to meet diagnostic criteria on retest. We operationalized severity as a ratio of the sum of all endorsed “stem” questions (those asked of all respondents, regardless of the skip structure built into the DISC) across the five major diagnostic categories, divided by the total number of possible stem questions. As expected, in the community sample, stable cases (n=71, positive for diagnosis at both time points) endorsed an average of 43.7% of time 1 stem symptoms compared with 33.4% among attenuating cases (n=74) ($t=-4.96$, $P=.0001$). Similarly, in the clinical sample, stable cases (n=64) endorsed 53.4% of time 1 stem symptoms compared with 41.4% among attenuating cases (n=15) ($t=-3.04$, $P=.003$). Thus, attenuation at the diagnostic level is not simply a result of subjects switching to different symptom patterns at time 2 but reflects an absolute decrease in the number of symptoms.

Similarly, subjects in the community sample who had used mental health services in the past year (n=46) endorsed an average of 39.0% of time 1 stem symptoms compared with 28.7% among nonusers (n=232) ($t=-4.38$, $P=.0001$). Finally, we examined the proportion of community cases who used services and compared service users and nonusers on diagnostic attenuation (attenuators vs nonattenuators). Among service users, 20 (69%) of 29 cases showed diagnostic stability, while among nonusers, only 54 (47%) of 116 cases showed diagnostic stability (Fisher’s Exact Test, $P=.04$). These findings further indicate that the lower reliabilities in community samples are related to a combination of factors, including decreased symptom severity, the presence of threshold cases, and other possible factors we did not examine (eg, mental health attitudes, stigma, burden).

**MODELING UNRELIABILITY**

As a final examination of the sources of unreliability, we adapted methods described by Rubio-Stipec and colleagues to perform regression analyses using computed $\kappa$ values as the dependent variable. Because there were five diagnostic categories, two algorithms (we used DISC-P and DISC-C only, since DISC-PC is the combination of the first two), and two settings (clinical and community) ($5\times2\times2=20$), there were 20 total $\kappa$ values. In the regression analyses we entered the DISC diagnosis source (1 for child algorithm, 2 for parent algorithm), setting (1 for community, 2 for clinic), and simple ratios for the four sources of unreliability.

According to the rationale for the computation of these four ratios, under ideal circumstances and perfect reliability, the population consists of persons who are either “true cases” or “true noncases.” Those who meet criteria at both time points are part of the population of true cases, while those who do not meet criteria at either time point are part of the population of true noncases. With fallible instruments administered under human conditions, however, four permutations of test-retest responses are possible: time 1 positive and time 2 positive (cell a [+1+]), time 1 positive and time 2 negative (cell b [+1–]), time 1 negative and time 2 positive (cell c [−1+]), and time 1 negative and time 2 negative (cell d [−1–]).

With these four cells, the ratio that models the error represented by persons who are from the population of true cases but fail to meet diagnostic criteria on sec-
ond interview (for any of a variety of reasons) is $b/(a+b)$, representing “true case attenuation.” Second, a group of subjects with a tendency to impulsively overendorse symptoms in novel settings may meet caseness criteria at time 1 even though they are actually part of the population of true noncases (cell d). The ratio to approximate this error is $b/(b+d)$, called for descriptive purposes “true noncase attenuation.” Third, some true noncases might endorse and meet criteria at time 2, possibly because of measurement error, random responding, etc. The ratio to approximate this potential source of information discrepancy is $c/(c+d)$, called for descriptive purposes “true noncase augmentation.” Fourth, some subjects who are part of the population of true cases could be slow to warm up, requiring a second interview before they are fully able to disclose their symptoms (or they may have developed the disorder in the test-retest interim). The ratio to model this “error” is $c/(a+c)$, called “true case augmentation.” Of course, any number of hypotheses might be invoked to explain each of these four sources of unreliability; we have only described several of the possibilities.

These four specific ratios were entered into the regression equation along with the informant source variable (child or parent) and setting variable (clinic or community). Stepwise and hierarchical regression analyses were performed to determine the percentage variance accounted for in the $\kappa$ statistics by these four sources of “error” and by the source and setting variables. As seen in Table 5, the regression was highly significant (as it should be, since all the variance in the $\kappa$ coefficients is represented by the four ratios), and the great preponderance of decreased $\kappa$ values was explained by the “true case attenuation” ratio (usually more than four times that of the other variables). A number of related analyses were run, systematically forcing each of the other variables and potential sources of unreliability first into the model; the results of these analyses were similar, in that the “true case attenuation” ratio continued to account for over 80% of the variance in $\kappa$ values. The informant and setting variables did not enter into any models, as all the variance was all accounted for by three of the four ratios. Other terms did not enter into the final model, so no $\beta$ weights or $t$ values are noted in Table 5 for them (more extensive information is available from the authors on request).

**Table 4. Test-Retest Reliability for All Sites Using the Parent vs Child Diagnostic Algorithm**

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Clinic Cases</th>
<th>Community Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parent (N)</td>
<td>Child (N)</td>
</tr>
<tr>
<td>Any anxiety disorder</td>
<td>58 (97)</td>
<td>20 (278)</td>
</tr>
<tr>
<td>Depression and/or dysthymia</td>
<td>69 (96)</td>
<td>00 (277)</td>
</tr>
<tr>
<td>Attention-deficit hyperactivity disorder</td>
<td>69 (88)</td>
<td>57 (265)</td>
</tr>
<tr>
<td>Oppositional defiant disorder</td>
<td>67 (95)</td>
<td>65 (274)</td>
</tr>
<tr>
<td>Conduct disorder</td>
<td>70 (88)</td>
<td>66 (278)</td>
</tr>
</tbody>
</table>

**COMMENT**

Before we comment on our findings, several caveats are necessary. First, there were major differences among sites in sampling procedures and subject characteristics as well as variations in ethnicity, culture, and language. Although these differences were part of the planned, allowable site variations during the first phase of the MECA Study, they nonetheless constitute important methodologic differences and possible study limitations. Regardless, the examination of site-specific data revealed only minor variations across sites. In fact, the patterns of reliability (eg, clinic vs community and parent vs child) remained fairly stable across all sites, particularly in diagnoses for sites with at least five time 1 cases.

Furthermore, because most clinical subjects were male (whereas community subjects were equally distributed across gender) and because of the longer test-retest intervals in the community subjects, it cannot be determined whether the higher clinical test-retest reliabilities are best explained by clinical referral, male gender, or the length of the test-retest interval. However, only one difference in reliability was found as a function of gender (depression and/or dysthymia for females vs males in the community sample only), which tends to rule out gender as a major confounder in the overall pattern of results. Similarly, differences in clinical vs community test-retest intervals (2 vs 3 weeks) were small and seem unlikely to explain the clinic-community reliability differences. In contrast, post hoc analyses of community subjects indicated that a history of mental health services use was related to increased diagnostic stability, as was increased symptom severity. In our view, both of these factors are more closely related to our observed clinic-community reliability differences.

**COMPARISONS WITH PREVIOUS STUDIES**

In general, our reliability findings are consistent with or superior to those reported in previous studies. Compared with those found using earlier versions of the DISC studied in clinical samples, our clinical sample reliabilities are substantially better for CD, moderately better for ADHD, and equivalent or better for depressive disorders. Similarly, we found superior reliabilities for test-retest reliabilities of two anxiety disorders studied in earlier versions of the DISC. For separation anxiety disorder and .57 for overanxious disorder (further details of specific anxiety and depressive disorder reliabilities are available on request from the authors). Compared with test-retest studies of other diagnostic instruments, clinical sample reliabilities reported here are comparable or superior, despite the need for clinically trained interviewers for most other instruments.

Unfortunately, there are few points of comparison for our community sample reliability findings. Two recent reports used a two-stage sampling design to select community subjects with a high probability of diagno-
sis at the second stage. However, the two-stage design of these studies makes direct comparisons of $\kappa$ statistics with our data difficult, since such sampling strategies increase disorder base rates and sample heterogeneity and result in higher $\kappa$ coefficients in equivalent (ie, fixed levels of sensitivity and specificity) diagnostic instruments. Nonetheless, these investigators reported much lower community sample reliabilities than in previous test-retest studies in clinical samples using the same instrument. Our community sample test-retest reliabilities do compare favorably (generally superior) with the reliabilities of the DIS used in the adult Epidemiologic Catchment Area studies despite the challenges entailed in obtaining and combining information from two informants in the MECA Study. However, DIS reliability data may not be directly comparable to our results, since the DIS test-retest study was conducted with two different types of interviewers (lay interviewers at time 1, clinicians at time 2).

### ADEQUACY OF THE DISC

Given the lower test-retest reliability of the DISC in the community, are lay interviewers are really "up to the task" of collecting complex diagnostic data for epidemiologic studies? The $\kappa$ coefficients reported in the present study must be considered in light of the fact that an instrument with acceptable characteristics (eg, sensitivity, specificity, test-retest $\kappa$ coefficients) in clinical settings will demonstrate lower test-retest $\kappa$ coefficients in epidemiologic (homogeneous) samples with low base rates. As others have noted, this does not indicate a problem with $\kappa$ statistics per se but reflects the true level of difficulty in obtaining agreement between two raters of less common conditions in the community. We suggest that in community samples with low disorder prevalence, "acceptable" $\kappa$ values may be overly conservative, and the lower community sample reliability values reported here for the internalizing disorders (26 to .32) should not be dismissed as too unreliable for use in epidemiologic studies.

Could clinical interviewers do better? In a recent head-to-head comparison with the DISC, clinician-generated diagnoses proved less reliable than lay interviewer-administered DISC diagnoses. Large-scale epidemiologic studies with clinical interviewers are probably not logistically feasible, and clinician-generated diagnoses are likely to be quite unreliable in community settings.

It is unfortunate that our samples only included children age 9 years and older, given the relatively large proportion of younger children who come to clinical attention and are in need of services. While previous research has indicated that problems with test-retest reliability increase as a function of younger age of the child informant, this is not necessarily the case with parental informants. In fact, evidence suggests that some diagnoses may be more reliable with the parents of younger children than with the parents of older children. Certainly, more information is needed concerning the psychometric properties of diagnostic interviews for younger children.

Given the greater reliabilities among parents than their children, could one use the parent report alone for diagnostic purposes? Such a strategy is problematic because parents and children provide nonredundant information. For example, in the present report, in the community sample, most of the conduct, anxiety, and affective disorders were identified through information obtained from the DISC-C. While opposite findings were seen in the clinical sample (most cases were identified through the DISC-P), increasing evidence suggests that child-derived diagnoses in community subjects are reliable and, more important, have long-term prognostic significance. Instead of surrendering children as potential informants, increased research is needed to determine under which circumstances (eg, clinical vs community settings) which diagnoses are best determined from which informants (parent, child, or both) for which children (eg, as a function of age and gender). Certainly, it would be dubious to fail to obtain information directly from children for internalizing disorders or conduct disorder, where the child often does not reveal symptoms to the parent. Even in the absence of parental endorsement, the child’s report of symptoms may have meaning in terms of subthreshold conditions, other diagnostic entities, or risk for future disorder.

### ATTENUATION

Collectively, our analyses indicate that test-retest changes in DISC-based diagnostic status are not just a function...
of random error but are significantly more likely to occur around the diagnostic threshold, occur more commonly in community than in clinic cases, and are related to symptom severity at time 1. In addition, for specific diagnoses, decreased reliability may be related to children's age (depression, CD) and gender (depression).

The level of test-retest attenuation we report here is not unique to childhood diagnostic interviews but is also a significant problem for adult diagnostic interviews, including semistructured interviews conducted by clinicians and highly structured interviews conducted by lay interviewers. This perservativeness of test-retest attenuation has led some investigators to argue that diagnostic instability should be regarded as a phenomenon worthy of study in its own right, with particular attention to personal and situational factors that are systematically related to test-retest differences in symptom reporting. It is probably too narrow a perspective to dismiss such phenomena simply as "unreliability," since systematic, measurable (and perhaps alterable) sources of variance (heretofore presumed to be "error") underlie significant amounts of test-retest attenuation.

Although test-retest attenuation in psychopathology research very much reflects a phenomenon in need of an explanation, its perservativeness across measures, methods, and informants suggests the possibility that traditional retest strategies in this domain may constitute a poor assay of a measure's reliability. As Robins has suggested, the diagnostic interview process itself appears to change subjects' response sets in ways that make a closely spaced retest interview difficult to interpret. If such a reactivity effect is operating, the comparison of two interviews over a relatively short time may be no more interpretable than retesting the effects of an acute psychological stress on heart rate immediately following an initial test. In both cases, it may be necessary to wait until a subject's reactivity to the initial stimulus condition has returned to its baseline level.

Any number of mechanisms may underlie interview-triggered changes in subjects' response sets. One possibility is that the intensive nature of diagnostic interview questions alters an individual's threshold for symptom reporting. The mere act of reporting a symptom as present may have a cathartic effect on some individuals, so that they later de-emphasize the significance of those symptoms. This hypothesis has not been examined in the literature, but it is reasonable to predict that any cathartic effect on symptom reporting would be short-lived, with subjects returning to their baseline evaluative levels following a sufficient "washout" period. Existing data (including those reported above) suggest that the catharsis hypothesis will need to be tested with test-retest durations exceeding the standard retest interval of 2 weeks.

Another source of test-retest attenuation may be a subject's conscious or unconscious desire to shorten the second interview by saying "no" to more symptom questions, having learned that "yes" responses during the initial interview result in additional questions and a longer interview. To the best of our knowledge, this hypothesis has not been tested in either the adult or child psychopathology literature. An optimal test of this hypothesis requires a design with symptom questions presented in counterbalanced order on two occasions. Support for the endorsement-avoidance hypothesis would be consistent with a significant association between likelihood of symptom endorsement and the temporal ordering of symptoms within either interview or with a negative association between the length of the first and second interviews. This hypothesis cannot be tested on our data because symptom questions were presented in a fixed order, with more severe diagnoses (and therefore lower probabilities of symptom endorsement) coming toward the end of the interview. Interestingly, Ribera and colleagues (J. Ribera, PhD, G.C., M.R.-S., et al, unpublished results, 1992) have noted that the test-retest attenuation may be lower when the subsequent interview is conducted by a clinician, and they note the possibility that the interviewee continues to tell the complete story on retest, possibly because he or she may feel that the physician would notice the discrepancy between the first report and the subsequent one. These alternative explanations will be examined in more detail in subsequent reports.

The post hoc regression analyses do not in any sense "prove" which if any of the cells constitute the "true" cases (eg, +/+ +/−, /+/− or /+/+). However, because psychopathologic conditions are not usually seen as socially desirable, because of the potential cathartic effects of the first interview, and because subjects may wish to shorten the second interview by saying no (having learned that "yes" responses result in additional questions and a lengthier interview), we tend to place greater credence on the veracity of the first interview (+/+ +/− subjects) rather than on the smaller number of cases who met the criteria for diagnosis in both time periods (+/+, subjects only). Our analyses of the four ratios of potential unreliability (Table 5) tend to support this position, but further studies specifically examining this question are needed.

Attenuation appears to be higher in children than adults, possibly because children may be more vulnerable to cognitive distortions of the meaning of the retest. For example, children may fail to understand the meaning of the retest, thinking that they need not repeat information given earlier (assuming it is now known), or they may conclude that the interview questions are being repeated because their previous responses were incorrect. If such explanations are accurate, the actual nature of the preparation of subjects for test-retest designs is of major importance for future studies.

Other factors that may explain parent-child attenuation and reliability differences include the possibility that children are more likely to forget the details of the previous interview or attach less significance to the rereporting of all symptoms or to the diagnostic process, so that they discount symptoms more readily (resulting in greater attenuation on the second interview). Also, children are likely more impatient than adults, have a shorter attention span, and have a lower tolerance for boredom. If so, a second interview will be less novel than the first and...
may fail to hold their attention, with the result that they may tend to deny symptoms the second time around, attempting to rush through the interview.

A number of issues particular to child and adolescent psychopathology warrant special mention. For example, the relatively greater reliability of externalizing disorders (ADHD, ODD, and CD) than internalizing disorders (depression and/or dysthymia, any anxiety disorder) (Table 4) suggests that externalizing disorders are more “noticeable” and problematic to parents (hence remembered and recalled at time 2), while internalizing disorders are more subjective, transient, and prone to recall difficulties. Furthermore, much of the parent’s and child’s recall (as in studies of adult psychopathologic conditions and life stress research) may be colored by the reporter’s current emotional state.32,33 We have relatively little information about such questions, and systematic studies of these phenomena are needed.

**CONCLUSIONS**

In summary, the test-retest reliability data from the MECA Study range from moderate to substantial39 in clinical settings and are comparable or superior to the reliabilities reported for other child and adolescent diagnostic instruments, despite the stringent conditions to which the DISC was submitted. However, the test-retest reliabilities from the community samples range from only fair (internalizing disorders) to substantial (externalizing disorders).34

Is the DISC too “unreliable” for use in field surveys? Given the ubiquitous difficulties inherent in test-retest designs (particularly in community samples with low base rates, decreased diagnostic severity, and increased attenuation), we suggest that such questions are better phrased as, “What is the validity of diagnostic information obtained at a single (first) time point?” and “Which strategies will decrease diagnostic attenuation, particularly in longitudinal studies that require repeated diagnostic assessments?” Further analyses from the second phase of the MECA studies will examine the validity of DISC diagnoses with respect to clinician interviews and external impairment criteria (eg, impairment, need for and use of mental health services). If, as expected, these data indicate robust validity correlates of time 1 DISC interviews, the DISC should provide meaningful prevalence estimates in community surveys. However, further studies of diagnostic attenuation are needed to examine variations in the DISC by reporter (parent vs child), by interviewer (layperson vs clinician), and by diagnosis. Studies of the effects of child and parent age, intelligence, attributions about symptoms, and the context in which the child and/or parent describes the child’s symptoms are essential. Innovative strategies to reduce the effects of repeated assessment procedures on symptom reporting are very much needed.

Accepted for publication June 13, 1994.

The opinions and assertions contained in this article are the private views of the authors and are not to be construed as official or as reflecting the views of the Depart-ment of Health and Human Services or the National Institute of Mental Health.

The logic for the computer algorithms used to process the responses was developed by a team composed of David Shaffer, MD, Prudence Fisher, MS, and John Piacentini, PhD. The computer programming was done by Mary Rojas, PhD, Michael Parides, MS, and Mark Davies, MPH.

The MECA Program is an epidemiologic 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 National Institute of Mental Health (NIMH), Rockville, Md. The NIMH Principal Collaborators are Darryl A. Regier, MD, MPH, Ben Z. Locke, MSPH, Peter S. Jensen, MD, William E. Narrow, MD, MPH, and Donald S. Rae, MA; the NIMH Project Officer was William J. Huber. The Principal Investigators and Coinvestigators from the four sites are as follows: Emory University, Atlanta, Ga (U01 MH46725): Mina K. Dulcan, MD, Benjamin B. Lahey, PhD, Donna J. Brogan, PhD, Sherryl Goodman, PhD, and Elaine Flagg, PhD; Research Foundation for Mental Hygiene at New York State Psychiatric Institute, Columbia University, New York (U01 MH46718): Hector R. Bird, MD, David Shaffer, MD, Myrna Weissman, PhD, Patricia Cohen, PhD, Denise Kandel, PhD, Christina Hoven, PhD, Mark Davies, MPH, Madelyn S. Gould, PhD, and Agnes Whittaker, MD; Yale University, New Haven, Conn (U01 MH46717): Mary Schwab-Stone, MD, Philip J. Leaf, PhD, Sarah Horwitz, PhD, and Judith H. Lichtman, MPH; and University of Puerto Rico, San Juan (U01 MH46732): Glorisa Canino, PhD, Maritza Rubio-Stipe, MA, Milagros Bravo, PhD, Margarita Alegria, PhD, Julio Ribera, PhD, Sarah Huertas, MD, and Michael Woodbury, MD.

Reprint requests to Room 18C-17, Parklawn Bldg, Child and Adolescent Disorders Research Branch, National Institute of Mental Health, 5600 Fishers Ln, Rockville, MD 20857 (Dr Jensen).

**REFERENCES**


