Psychological Medicine, 1995, **25**, 755–762. Copyright © 1995 Cambridge University Press

A test–retest reliability study of child-reported psychiatric symptoms and diagnoses using the Child and Adolescent Psychiatric Assessment (CAPA-C)

A. ANGOLD[1] AND E. J. COSTELLO

From the Developmental Epidemiology Program, Department of Psychiatry, Duke University Medical Center, Durham, NC, USA

SYNOPSIS Seventy-seven 10–18-year-old psychiatric in-patients and out-patients took part in a test–retest study of the Child and Adolescent Psychiatric Assessment (CAPA). They were interviewed on two occasions several days apart. Overall reliability of diagnosis ranged from $\kappa = 0{\cdot}55$ (conduct disorder) to 1·0 (substance abuse or dependence). In general, reliability for scale scores of psychopathology was somewhat lower in out-patients than in-patients, though the opposite was the case for anxiety disorders and psychosocial incapacity and the reliability of the diagnosis of conduct disorder – the only individual diagnosis sufficiently common to permit this comparison. Unreliability of reports of behavioural problems was found to be related to admitting to being a liar in the first interview. The implications of these results for the use of the CAPA are discussed.

INTRODUCTION

In the preceding paper (Angold *et al.* 1995) a discussion of the advantages and disadvantages of interviewer-based and respondent-based approaches to diagnostic interviewing in psychiatry, and a description of the Child and Adolescent Psychiatric Assessment (CAPA) (Angold *et al.* 1992) were presented. The CAPA is a diagnostic interview for children and adolescents (aged 8–18) that combines methods developed in both the interviewer- and respondent-based traditions. Here we present the results of a test–retest study of CAPA interviews conducted with children and adolescents, and discuss their implications for its use in psychiatric research and clinical practice.

METHOD

Subjects

The subjects of this study were a consecutive sample of 77 10–18-year-olds attending for clinical services at a day hospital, two in-patient units, or a child guidance clinic in Durham, North Carolina. Subjects involved in other studies were excluded because of the substantial interview load involved, but otherwise, the only exclusion criterion was an IQ below 70. No child would have been included had the clinician in charge of the case felt that the procedure would be detrimental to his or her state. However, in no case was this condition invoked. Table 1 shows some characteristics of the children in the sample: 57% of the sample was African American; 7% were of mixed race and the remainder (36%) were white.

Subjects from the private day hospital and private in-patient units tended to be of higher socioeconomic status than the public out-patient clinic subjects, who were mostly either uninsured or reliant on Medicaid insurance.

Table 1. *Characteristics of the sample*

	N	%
Gender		
Male	45	58
Female	32	42
Age and clinical status		
10–13 years in-patient	15	19
10–13 years out-patient	17	22
14–18 years in-patient	24	31
14–18 years out-patient	21	28

[1] Address for correspondence: Dr Adrian Angold, Developmental Epidemiology Program, Department of Psychiatry, Duke University Medical Center, Box 3454, Durham, NC 27710, USA.

Informed consent for participation in the study was obtained both from the child and a parent or guardian, and, in most cases, both the parent and child were interviewed. A payment of $10 was made to subjects when they completed the protocol.

Interviewers

The interviews were conducted by a graduate team of bachelor ($N = 12$) or master level interviewers ($N = 2$) who had received 1 month of initial training, and were also conducting CAPA interviews as part of a separate study of co-morbidity between substance abuse and psychiatric disorders in adolescents. Fourteen interviewers completed between one and 28 interviews each.

Procedure

Once informed consent had been obtained, two interviews were scheduled to take place either at the subject's home or at the study offices. Subjects were free to choose whichever location they preferred. The two interviews were initially scheduled to take place within 1 week of each other. The final time between interviews was between 1 and 11 days (mean = 5 days). Different interviewers were used for each interview. All interviewers were blind to any clinical or CAPA interview information about the subjects at the start of their interviews. Care was taken to ensure that other family members could not overhear any part of the interview. Subjects were assured that all information was confidential, and would not be passed on to other family members.

The CAPA was administered as part of a package of related instruments, but it constituted the single largest item in that package. The other instruments dealt with family functioning, mental health service use, and symptom scores. An upper limit of 3 h was placed on the total duration of the entire interview package, in order to keep subject burden within acceptable limits. In 16 cases this meant that the CAPA was not completed in its entirety. For diagnoses where the relevant information for making a diagnosis was not collected in one or other of the interviews, those cases were excluded from the analyses of reliability for that diagnosis.

The interviews were tape-recorded, unless the subject specifically requested that they should not be taped. The interviewer also kept detailed notes throughout the interview, in order to facilitate later coding in the office. The total interview package (including the CAPA and other instruments) with children took between 1·5 and 3 h, depending on the number of symptoms and the child's loquacity. The median length was 2 h. The CAPA itself consumed on average about 1 h. All interviews were checked before data entry by a second interviewer or interviewer supervisor (usually the latter) with extensive CAPA interviewing experience, and doubtful codings were discussed with the interviewer to determine the proper coding. For 20 of the 77 interview pairs, both interviews were checked by the same person.

Data analysis

Diagnosis, symptom scores and scale scores were generated from the raw interview symptom data by a series of computer algorithms (the CAPA-Originated Diagnostic Algorithms, CODA). For the purposes of comparison with other structured interviews, we report only on DSM-III-R diagnoses and symptom scales here. In accordance with current common practice, the comorbidity exclusion criteria (such as not making the diagnosis of dysthymia in the presence of a major depressive episode) were ignored.

Cohen's κ (Cohen, 1960) was used to assess agreement on categorical variables (diagnoses) while scale score agreement was measured by the intraclass correlation coefficient (ICC).

RESULTS

Diagnostic reliability

Table 2 shows levels of agreement on those diagnoses that occurred with sufficient frequency to warrant κ analyses (as least five cases reported by one interviewer or the other). It can be seen that diagnostic agreement ranged, using the terminology of Landis & Koch (1977) from moderate (conduct disorder) through good (dysthymia, psychosocial incapacity) to excellent (overanxious disorder, major depressive disorder, substance abuse/dependence). Since only one child met criteria for self-reported oppositional defiant disorder, the intraclass correlation for the item scores is presented. In the case of obsessive compulsive disorder, only

Table 2. *Agreement above chance between two interviewers*

Diagnosis (number of subjects available for analysis)	a*	b*	c*	d*	κ	P_{pos}	P_{neg}	OP ICC (Mean for OP)	IP ICC (Mean for IP)	Total ICC (*N*)
Major depression (77)	71	1	0	5	0·90	0·91	0·99	0·58 (0·72)	0·91 (1·93)	0·88 (77)
Dysthymia (77)	73	1	0	3	0·85	0·86	0·99	0·44 (0·85)	0·78 (1·42)	0·69 (77)
Depression or dysthymia (77)	70	2	0	5	0·82	0·83	0·99	0·51 (1·05)	0·90 (2·54)	0·86 (77)
Over-anxious disorder (74)	69	2	0	3	0·74	0·75	0·99	0·87 (1·03)	0·73 (1·41)	0·80 (77)
Separation anxiety disorder (74)	—	—	—	—	—	—	—	0·47 (0·29)	0·79 (0·43)	0·63 (77)
Generalized anxiety disorder (74)	71	0	1	2	0·79	0·80	0·99	—	—	—
Any anxiety disorder (74)	66	3	1	4	0·64	0·66	0·97	0·82 (0·69)	0·73 (1·47)	0·78 (77)
Conduct disorder (70)	54	3	6	7	0·55	0·61	0·94	0·45 (0·78)	0·66 (1·56)	0·60 (70)
Oppositional disorder (70)								0·37 (0·79)	0·66 (1·38)	0·50 (70)
CD or ODD (70)	53	3	7	7	0·50	0·58	0·91	0·45 (1·57)	0·75 (2·74)	0·61 (70)
Substance abuse/dependence (68)	65	0	0	3	1	1	1	— (0)	0·98 (0·53)	0·98 (61)
Psychosocial incapacity (65)	—	—	—	—	—	—	—	0·77 (2·66)	0·73 (5·98)	0·76 (66)

κ = Kappa (Cohen, 1966).

P_{pos}, proportion of positive agreement; P_{neg}, proportion of negative agreement; ICC, intraclass correlation; please see text for explanation. OP, out-patient; IP, in-patient.

a*, Negative at both interviews; b*, Negative at first interview and positive at second; c*, Positive at first interview and negative at second; d*, Positive at both interviews.

two symptom ratings go into the DSM-III-R diagnostic criteria, so an ICC is not useful. Similarly, the way in which the symptoms of generalized anxiety disorder are collected do not lend themselves to the construction of a scale score, so only diagnostic reliability is reported for generalized anxiety disorder. Psychosocial incapacity is a scale score made up of all incapacity ratings; here too the intraclass correlation for item ratings was used as the index of reliability.

Reliability of symptom scale scores

Scale scores representing counts of DSM-III-R symptoms for various diagnostic categories were constructed from the conduct, oppositional, depression, anxiety, substance abuse and psychosocial incapacity sections of the CAPA. These symptoms counts included the DSM-III-R symptom frequency and duration ratings. Intraclass correlation coefficients for each of these scales are also shown in Table 2. Since it has been suggested (Jensen *et al.* 1992) that reliability is, at least in part, a function of the severity of diagnosis (with reliability being higher with more severely affected groups), we also computed ICCs separately for the in-patient and out-patient subsamples. The in-patient sample had substantially higher scores on all of the scale measures, but the differences in levels of reliability between the two groups varied in direction. For major depression, dysthymia, conduct disorder and oppositional disorder, the ICCs were higher in the in-patient group. On the other hand, out-patient reliabilities were higher for the anxiety disorders and psychosocial incapacity. There were enough cases of conduct disorder to allow the κ values to be calculated separately for each of these groups, and there was essentially no difference in the reliability of this diagnosis between in-patients ($\kappa = 0{\cdot}50$) and out-patients ($\kappa = 0{\cdot}53$).

Symptom attenuation as an explanation for unreliability

While these results are satisfactory when compared with the diagnostic reliability of other instruments (Herjanic & Campbell, 1977; Hodges *et al.* 1981, 1982; Herjanic & Reich, 1982; Chambers *et al.* 1985; Costello *et al.* 1985), we were concerned to understand better the reasons for unreliability. Unreliability with the Diagnostic Interview Schedule for Children (DISC) (Shaffer *et al.* 1989), for example, appears to be substantially attributable to what has been called the 'attenuation effect' (Costello *et al.* 1984; Jensen *et al.* 1992). Children report lower rates of disorder on the second interview than they did on the first interview. We, therefore, began by examining our data for such attenuation. Since this was a relatively small study compared with the DISC studies, simply looking at rates of diagnosis proved unsatisfactory. However, the scale scores allowed us to examine this question for various symptom areas. Table 3 shows that the mean symptom scores did indeed fall at the

Table 3. *Mean symptom scores on two occasions of measurement*

Symptom area	Mean score at interview 1 (50)	Mean score at interview 2 (50)	% variance explained by occasion of measurement	*P*
Depression	1·33 (2·23)	1·23 (2·13)	0·05	0·77
Anxiety	1·05 (1·79)	0·77 (1·75)	0·7	0·32
Conduct	1·34 (1·49)	0·89 (1·23)	2·7	0·05
Oppositional	1·31 (1·56)	1·03 (1·49)	0·9	0·27
Substance abuse	0·24 (1·11)	0·23 (0·97)	0	0·93
Psychosocial incapacity	4·21 (4·79)	4·03 (5·31)	0	0·83

second interview, but that, except in the case of conduct disorder symptoms, these reductions were rather small (with the means differing by an average of around one-tenth of a standard deviation). In the case of conduct disorder the 'attenuation' was at a level of around one-third of a standard deviation.

Lying and unreliability

Lying might well be expected to result in unreliability. The CAPA-C (child interview) includes self-reports of the number of lies told during the preceding 3 months, and so we compared the number of 'attenuators' (subjects with lower scale scores at the second interview) among those who reported in the first interview that they had told at least two lies with the number of 'attenuators' among those who said that they did not tell lies. While lying was not associated with the pattern of unreliability in relation to depressive or anxiety symptoms, it was substantially associated with the pattern of unreliability in reports of conduct symptoms. Seventy-four per cent (20 of 27) of those who admitted to lying in the first interview were 'attenuators', while only 42% of those who did not report being liars were 'attenuators' ($P = 0{\cdot}013$ from two-tailed Fisher exact test). These findings suggest that the attenuation effect is really present in relation to conduct disorder symptoms. We are, therefore, seeing a systematic difference between interviews in the area of reports of conduct problems, and one conclusion must be that the test–retest paradigm can induce a change in subjects' reporting characteristics, particularly in those who are prone to lying.

This pattern of responses helps in deciding which report is the one to believe. Those whose reports on lying differ from one occasion to another must be telling other than the truth on one of these occasions. In other words they must be liars, at least on the subject of lying. We may, therefore, conclude that they are being more truthful on the occasion on which they admit to lying. Assuming that this increased level of truthfulness is a characteristic of the whole interview, and not just the lying item, we would then conclude that the occasion on which more symptoms are reported (since it is associated with truthfulness about lying) represents a more accurate representation of the child's behaviour. This piece of logical sophistry leads to the same conclusion as the common sense clinical approach, which indicates that it is unlikely that many individuals fabricate stories of behavioural deviance. The interviewers' notes and tapes, which often contained detailed descriptions of deviant behaviours, lend further weight to the idea that these convincing stories were unlikely to have been invented on the spur of the moment. In other symptom areas, however, we have little evidence that unreliability in CAPA interviews is due to other than random variation.

Impact of interview checker non-blindness on reliability

A potential problem for this study is the fact that 20 of the 77 interview pairs had the same checker who looked at both interviews for coding errors. This situation arose when no other checker was available to complete this work in a timely manner. This non-blindness could have been responsible for elevations in the level of agreement between interviewers. We allowed this situation to arise because we did not expect it to make any difference. A further check on coding integrity consisted of weekly meetings between the interviewers, checkers and the first author at which interviewer's tapes and codings were reviewed and critiqued for quality control purposes. Thus, any 'massaging' of the codings resulting from a checker having seen a previous interview schedule with a subject would have been likely to be detected.

Table 4. *Comparison of ICCs for interviews checked by different* v. *same checker*

	Blind checker (ICC)	Non-blind checker (ICC)
Depression	0·82	0·88
Anxiety	0·80	0·81
Conduct	0·62	0·50
Oppositional	0·61	0
Substance abuse/dependence	0·98	1·0
Psychosocial incapacity	0·74	0·85

Table 5. *Agreement on the amount of time (duration or frequency) that a symptom was present during the past 3 months*

	N reporting symptom	ICC
School time missed due to worry/anxiety	35	0·90
Anticipatory fear of school	15	0·53
Separation worries/anxiety	17	0·37
Worries	46	0·47
Nervous tension	19	0·49
Depressed mood	49	0·75
Anergia	16	0·77
Irritability	23	0·62
Rule breaking	27	0·94
Disobedience	29	0·30
Lying	72	0·37
Cheating	15	0·43
Forgery	16	0·29
Fights	43	0·76
Amount smoked	44	0·59
Amount of alcohol drunk	30	0·95

In order to assess whether checking had resulted in elevations of reliability coefficients, we conducted separate analyses of blindly-checked and non-blindly-checked interview scale scores. These results are presented in Table 4. It can be seen that for two of the six scales, agreement was actually lower in the non-blind group, while for the remaining four scales it was slightly higher. The two largest differences were both in the direction of the non-blind interviews having lower ICCs. We conclude, therefore, that the inclusion of non-blindly checked interviews did not inflate the reliability estimates.

Reliability of individual symptom ratings

A major feature of the CAPA is that it collects duration and frequency estimates for a number of emotional symptoms. When symptom duration and frequency are multiplied together, the result is an estimate of the amount of time during the preceding 3 months that the symptom has been present. For many other symptoms (particularly in the sections on oppositional and conduct problems where duration is not a relevant dimension), information on the frequency of problems is collected. Table 5 presents ICCs for such ratings of individual symptoms for those items that were reported positively in at least 15 interviews. We include it here because many DSM diagnostic rules include frequency or duration criteria.

DISCUSSION

Before discussing the implications of these data for the use of the CAPA, and modifications we have made to the instrument in the light of them, we will first discuss two aspects of the study that may seem odd at first sight: the low rate of diagnosis, and the lack of diagnoses of oppositional defiant disorder.

Many subjects did not receive a diagnosis

One problem with strictly implementing a set of diagnostic rules, such as those embodied in the DSM-III-R, is that many clinically-referred individuals do not receive any diagnosis (Rutter & Shaffer, 1980), since they do not quite meet criteria for any individual diagnosis. Thus, the child who reports two conduct disorder symptoms, four symptoms of oppositional disorder and four depressive symptoms that have lasted 5 months does not meet criteria for any of these disorders. Such a child could be classified as suffering from 'Depressive Disorder Not Otherwise Specified', but since no positive inclusion criteria are specified for this diagnosis beyond the statement 'disorders with depressive features that do not meet the criteria for any specific Mood Disorder or Adjustment Disorder with Depressed Mood' (p. 233), almost any pattern of depressive symptoms could be included or excluded at the whim of the investigator. Thus, we concentrated here only on more clearly defined diagnoses with the result that many subjects did not receive any diagnosis. One advantage of this approach is that from the point of view of diagnostic reliability it represents a stringent test, since κ is sensitive to prevalence.

It is generally accepted that to make a final diagnosis, information from both parents and children is desirable. Many of these children would probably have received diagnoses on the basis of parental reports (particularly in relation to oppositional disorder and attention deficit/hyperactivity disorder). However, combining parent and child reports begs the question that is the main focus of this study; whether the child reports themselves can be relied upon.

Where have all the oppositional disorders gone?

At the start of this study, we expected oppositional disorder to emerge as one of the most common diagnoses, but in the end there was only a single case of oppositional disorder based on child reports. It is now clear that this was due to our having set an inappropriately high cut-point for coding oppositional behaviour for child reports. By and large, behaviours in this section were only coded if the child reported that they had occurred three or four times a week (depending on the symptom) for at least 6 months. In contrast, the DISC, for instance, requires only weekly occurrence of such symptoms. We had regarded only weekly defiance and the like as being too common to be an appropriate cut-point. However, it is now apparent to us that children will only rarely report such frequent oppositionality across a range of items. This approach is also at least partially responsible for the relatively poor reliability of the oppositional behaviour scale scores, since with a child who reported being touchy or easily annoyed 20 times a week for 6 months, that symptom was coded as being absent, and its frequency automatically defaulted to 0. If, on the second interview, the child reported the symptom as occurring 30 times a week for 6 months, the symptom frequency would have been coded as 30 × 12 (the number of weeks in the primary period of the interview) = 360. Clearly, our present cut-offs were mistaken. We have modified the items that were in this format so that we now determine whether the behaviour in question ever occurs, and then note its frequency and when it began.

Identifying causes of unreliability

The strictest test of the reliability of an instrument uses the test–retest paradigm, in which the assessment is repeated within a period over which the phenomena being measured can be expected to remain stable. In the case of a diagnostic interview this means having two interviewers conduct a detailed interview on two occasions. Thus, unreliability in the test–retest design may result from any or all of five sources; (1) random variation in subjects' responses between interviews; (2) systematic variation in subjects' responses between interviews; (3) random variation in interviewers' administration and coding of the interview; (4) systematic variation in the interviewers' administration and coding of the interview; and (5) real changes in the child's symptomatology between the first and second assessments. A well calibrated instrument will be sensitive to real changes in symptoms, while minimizing the chances for variation in response that do not accurately reflect variation in symptomatology. Consideration of these sources of variation suggests that they have very different implications for improvement in interview reliability.

Random variation in a child's responses or interviewers' administration between two assessments implies imprecision in measurement, and indicates that an improved interview and/or better interviewer training would be desirable. However, systematic variation in responses between the first and second interviews may represent a weakness in the test–retest design itself. If reports of symptoms differ systematically between assessments then it is reasonable to argue that the retest paradigm has induced the difference (provided other factors have been held constant). To draw a physical analogy, if the measurement of haemoglobin required that 2 litres of blood be drawn for each assessment, one might expect that the results of the second measurement would be substantially lower than those of the first. One could not argue from these results that the method of haemoglobin estimation was unreliable, since the process of measurement itself was responsible for the difference between the two measures. Common sense informs us that the first measurement offered a truer picture of the hapless subjects' pretest haemoglobin levels. Where a convincing explanation can be provided for systematic variation between assessments, and it is apparent which of the measures is closest to providing an accurate estimate of the subject's status, such variation should not be regarded as necessarily implying a problem with the instrument itself.

However, such differences should indeed be regarded as a form of unreliability when these two conditions are not met.

The results presented above suggest that the non-random 'attenuation' effect with conduct and oppositional problems in our data means that the test-retest design is a flawed test of the 'quality' of reports in this area. The lying data also suggests that reports from the first interview represent the more accurate picture of the child's behavioural status. This conclusion may not, however, apply to other instruments in which attenuation occurs. A variety of possible explanations of attenuation in terms of learning effects, responses to hasten the end of the interview, and reconsideration of previous answers have been put forward (Jensen *et al.* 1995) and any or all of these could also be operating in relation to other interviews as could the effects associated with lying that we have reported here. However, we would not recommend using the CAPA in substantive studies in which symptoms need to be reassessed less than 3 months apart. Of course, the CAPA, with its 3-month primary period was not designed for such a purpose.

Random variation in the interviewers' administration and coding of the interview may result from problems with interview design or poor interviewer training, and respondent-based interviews have had substantial success in reducing these factors to a minimum. Supervisory, methodological and statistical procedures are also available to identify and prevent systematic interviewer bias both between interviews and across interviews. So, while these represent real sources of unreliability, they can be quantified within the test–retest design and, hopefully, controlled in practice. Interviewer drift is a problem with all types of interviews, and perhaps more so with interviewer-based assessments. We recommend that all CAPA interviews be checked by a supervisor to improve consistency across interviewers, and that samples of interviewer tapes be critiqued in regular sessions attended by all interviewers.

If an interview identifies real changes in a subject's state between two administrations, common sense tells us it is doing a good job. In fact, sensitivity to change is usually regarded as being a hallmark of psychometric excellent. However, unless some means are available to determine whether such a change has taken place, it will reduce the apparent reliability of the instrument. A review of all item disagreements in this study suggested that change over the period of 1 week sometimes occurred, as in the case of a subject who stated quite clearly that he had developed suicidal ideation during the week between his two interviews. However, such changes were uncommon and not a major contributor to the overall pattern of results.

Recommendations for the use of the CAPA

The results reported here indicate that the CAPA-C shows reasonable levels of retest reliability, even at the level of individual symptoms. Our findings on the Incapacity section indicate that the approach to psychosocial impairment embodied in the CAPA offers a way of collecting detailed information in this under-studied area. The CAPA may be particularly suited to the study of children's reports of emotional disorders given its satisfactory reliabilities in these areas, and its detailed quantitative ratings of the duration and frequency of emotional symptoms.

The time taken to conduct and process an interview is an important consideration for any study or clinical assessment. Our experience from several thousand CAPA interviews is that, in a general population sample, the introduction and symptom review require an average of 40 min and that the Incapacity section takes an average of 10 min. In addition, the interviewer requires 45 min to code the interview, and a further 30 min are needed for checking by a supervisor. Data entry takes 25 min per entry (50 min for double entry). It should be noted that interview completion time is more variable on interviewer-based interviews like the CAPA than on a respondent-based interview, because the child has more freedom to respond in his or her own words. The interview is also likely to require longer with clinical samples, because of their higher levels of symptomatology.

At the present time, the CAPA is being used in several clinical and epidemiological studies in England and the USA, and as a tool for clinical evaluation. In all of these settings it is critical to establish mechanisms for ensuring continuing interviewer adherence to interview protocol, but we and others have now had substantial experience with establishing such mechanisms, and successfully implementing the CAPA as a diagnostic instrument in studies ranging in size from less than 100 to over 1000 interviewed subjects.

REFERENCES

Angold, A., Cox, A., Prendergast, M., Rutter, M. & Simonoff, E. (1992). The Child and Adolescent Psychiatric Assessment (CAPA). DUMC, Box 3454, Durham, NC 27710, USA. (Unpublished report.)

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

Chambers, W. J., Puig-Antich, J., Hirsch, M., Paez, P., Ambrosini, P. J., Tabrizi, M. A. & Davies, M. (1985). The assessment of affective disorders in children and adolescents by semistructured interview: test–retest reliability of the schedule for affective disorders and schizophrenia for school-age children, present episode version. *Archives of General Psychiatry* **42**, 696–702.

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

Costello, A. J., Edelbrock, C. S., Dulcan, M. K., Kalas, R. & Klaric, S. H. (1984). Report on the NIMH Diagnostic Interview Schedule for Children (DISC). (Unpublished report.)

Costello, E. J., Edelbrock, C. S. & Costello, A. J. (1985). Validity of the NIMH Diagnostic Interview Schedule for Children: a comparison between psychiatric and pediatric referrals. *Journal of Abnormal Child Psychology* **13**, 579–595.

Herjanic, B. & Campbell, W. (1977). Differentiating psychiatrically disturbed children on the basis of a structured interview. *Journal of Abnormal Child Psychology* **5**, 127–134.

Herjanic, B. & Reich, W. (1982). Development of a structured psychiatric interview for children: agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology* **10**, 307–324.

Hodges, K., Klein, J., Fitch, P., McKnew, D. & Cytryn, L. (1981). The Child Assessment Schedule. *Catalog Selected Documents Psychology* **11**, 56.

Hodges, K., Kline, J., Stern, L., Cytryn, L. & McKnew, D. (1982). The development of a child assessment interview for research and clinical use. *Journal of Abnormal Child Psychology* **10**, 173–189.

Jensen, P. S., Shaffer, D., Rae, D., Canino, G., Bird, H. R., Dulcan, M. K., Lahey, B. B., Rubio-Stipec, M., Goodman, S. & Richters, J. E. (1992). Attenuation of the Diagnostic Interview Schedule for Children (Disc 2.1): sex, age and IQ relationships. Paper presented at the 39th Annual Meeting of the AACAP, October, Washington, DC.

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

Landis, J. R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Rutter, M. & Shaffer, D. (1980). DSM-II: a step forward or back in terms of the classification of child psychiatric disorders? *Journal of the American Academy of Child Psychiatry* **19**, 371–394.

Shaffer, D., Fisher, P., Piacentini, J., Schwab-Stone, M. & Wicks, J. (1989). Diagnostic Interview Schedule for Children (DISC-2.1C) Child Version. Department of Child Psychiatry, NY State Psychiatric Institute, 722 West 168th St., NY 10032, USA. (Unpublished report, available from first author.)