J. Child Psychol. Psychiat. Vol. 39, No. 7, pp. 963–972. 1998
Cambridge University Press
© 1998 Association for Child Psychology and Psychiatry
Printed in Great Britain. All rights reserved
0021–9830/98 $15.00+0.00

Characteristics of the Subject and Interview Influencing the Test–Retest Reliability of the Diagnostic Interview for Children and Adolescents-Revised

Rosario Granero Pérez, Lourdes Ezpeleta Ascaso, José M. Doménech Massons, and Nuria de la Osa Chaparro

Universitat Autònoma de Barcelona, Spain

This paper reviews some of the characteristics of the informants as well as some of the attributes of the DICA-R interview that could influence the test–retest reliability in a sample of 109 psychiatric outpatients aged 7–17 years. Different regression models using reliability coefficients constructed from the kappa statistic were obtained. Of those characteristics evaluated in the children, a high level of psychological impairment proved to be significant when it came to predicting the lowest test–retest reliability of the answers; none of the subject-related characteristics were significant in the adolescent patient model. The attributes of the questions that proved to be significant when explaining the lower reliability obtained for the individual question in the children's model were the length of the questions (longest questions), the content (internalising), the presence of time concepts, comparison with the peer group, and the need to exercise judgement; in the adolescents' model, the significant attributes were found to be the internalising content, the presence of time concepts, evaluation concerning the impairment caused by the disorder, and the need to exercise judgement. In the group of children our results are in accordance with the original paper. Similar results were found with adolescents. These findings have implications for the development and revision of new interview schedules.

Keywords: Test–retest, reliability, content, judgement, time concepts, structured diagnostic interview, Diagnostic Interview for Children and Adolescents-Revised (DICA-R).

Abbreviations: CGAS: Children's Global Assessment Scale; DICA-R: Diagnostic Interview for Children and Adolescents-Revised; DISC: Diagnostic Interview Schedule for Children.

Introduction

For a long time, the psychological assessment of young children was based on a clinical interview with the parents (Loeber, Green, & Lahey, 1990; Rutter & Graham, 1968; Williams, McGee, Anderson, & Silva, 1989), the argument being that children find it very difficult to pay attention during the assessment process and do not have the ability to talk about their own behaviour and feelings or to describe the symptoms of their psychological disorders (Herjanic, Herjanic, Brown, & Wheatt, 1975; Schwab-Stone, Fallon, Briggs, & Crowther, 1994). In recent decades, however, increasing numbers of researchers have agreed that parents do not always provide more reliable information than other informants (such as the child's teachers or the child himself or herself, for example). Nowadays it is accepted that assessment of children and adolescents should include various attributes: the involvement of a number of different informants, the use of various techniques (interviews, direct observation, self-assessment reports, etc.); and that it should take into account the subject's level of development and obtain information concerning the cognitive ability of the child (La Greca & Stone, 1992).

This change in direction has made it necessary to develop instruments for direct assessment with the child, in order to obtain standardised measurements of the disorders and the degree to which they exist, particularly in the case of epidemiological studies (Werry, 1992; Young, O'Brien, Gutterman, & Cohen, 1987a, b). Chief among these instruments are the so-called structured interviews (Ezpeleta, 1995, 1996; Silverman & Kearny, 1992).

It has recently been postulated that the content of the structured interviews most frequently used in the assessment of psychopathology could seriously challenge the cognitive and verbal skills of younger children (Young et al., 1987b; Zahner, 1991). The most recent reliability studies have therefore attempted to determine the specific contribution of individual variables such as age and even gender. The majority of studies carried out on gender

Requests for reprints to: Rosario Granero Pérez, Universitat Autònoma de Barcelona, Departament de Psicobiologia i de Metodologia de les Ciencies de la Salut, Edifici B, 08193 Bellaterra (Barcelona) Spain (E-mail: roser.granero©uab.es).

have not given results that clearly point to this variable being a determining factor as regards reliability (Canino et al., 1987; Edelbrock, Costello, Dulcan, Conover, & Kalas, 1986; Rapee, Barrett, Dadds, & Evans, 1994; Reich, Herjanic, Welner, & Gandhy, 1982); age, however, has proved to be an important variable to which some of the conditions of the psychological assessment should be subordinated. Some of the test–retest studies gauging the contribution of age have led to contradictory results; there are authors who maintain that this variable has no effect on the degree of agreement (Verhulst, Althaus, & Berden, 1987; Weissman et al., 1987); nevertheless, the majority of studies agree in pointing out that reliability increases with the age of the children (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985; Schwab-Stone et al., 1994; Silverman & Eisen, 1992).

Apart from the biases present in the data due to the characteristics of young informants (basically, their age), few studies rule out a certain amount of disagreement due to other factors inherent in the assessment process, such as the attributes of the questions put to the children. Edelbrock et al. (1985) indicated that the use of more complex sentences could reduce the level of reliability in the answers. In a pilot study, Herjanic and Reich (1982) found that the items showing the greatest reliability were specific questions referring to behaviour and events that are easily made objective and easy for the children to understand, as well as those that refer to symptoms that cannot go unnoticed. However, all these elements, together with other attributes of the various items, have yet to be studied in depth. For example, we still do not know the specific effect of grammatical complexity when the interview involves young children, exactly how questions using abstract constructs (such as time or emotion) influence the reliability of information obtained in the various age groups, the influence of the children's ability to relate their own thoughts and behaviour, and their difficulty in recounting ideas and behaviour that liken them to or set them apart from their peers.

Whereas these studies provide a useful starting point, few have explored the relationships between the characteristics of younger informants, the attributes of specific questions, and the reliability of the answers. To our knowledge, the only study of this type to date is the one by Fallon and Schwab-Stone (1994), who analysed the influence of some characteristics of the children aged 6–12 years and some of the attributes of the questions included in the Diagnostic Interview Schedule for Children (DISC; Shaffer et al., 1993) in a test–retest design. The results of this study indicated that by controlling age, cognitive ability, and gender, the children were more reliable concerning observable behaviour and less reliable as informants when responding to questions including unspecific time concepts, which involve reasoning about their own thoughts, or which require them to compare their behaviour with that of other children. As might be expected, parents' reports of their children were more reliable than their children's reports.

While we consider the previous research useful for the psychological assessment of children, the present work continues along these lines. In particular, we have replicated the Fallon and Schwab-Stone (1994) study, drawing upon a sample of Spanish-speaking psychiatric outpatients aged 7–17 years, using the Diagnostic Interview for Children and Adolescents-Revised (DICA-R; Reich, Shayka, & Taibleson, 1991). We have explored a number of characteristics relating to the informants (children and adolescents) and some of the attributes of the questions that could influence the test–retest reliability of the DICA-R. Three characteristics of the subjects are examined (gender, age, and degree of psychological impairment), as well as eight attributes of the questions (length, content, time concepts, clustering of symptoms, frequency, intensity of the disorder, peer comparison, and judgement).

Method

Participants

The sample comprises 57 children aged 7–12 years and 52 adolescents aged 13–17 years, who attended the psychiatric outpatients' clinic of the public hospitals network. For the children, 61 % of the data was for male patients, while 39 % was for female patients; for adolescents, 29 % of the data was for males, while 71 % was for females. Among the children, 98 % of the families were Caucasian and 2 % were Gypsy; the adolescent patients belonged 100 % to Caucasian families. Those subjects who were believed or known to be mentally retarded were excluded from the study.

Materials

The DICA-R is a semi-structured diagnostic interview that follows the criteria of the DSM-III-R (American Psychiatric Association, 1987). The DICA-R evaluates symptoms throughout life. There are three versions that are almost identical as regards structure and content: DICA-C (for children aged 6–12 years), DICA-A (for adolescents aged 13–17 years), and DICA-P (for parents). The present study used versions DICA-R-C and DICA-R-A (version 7.2).

On average, agreement between parents and children over the DICA answers was moderate (Herjanic et al., 1975; Herjanic & Reich, 1982; Kashani, Orvaschel, Burk, & Reid, 1985; Reich et al., 1982; Sylvester, Hyde, & Reichler, 1987; Welner, Reich, Herjanic, Jung, & Amado, 1987), a better level of agreement being obtained on observable disorders than on internal conditions. The first psychometric data from version 7.2 applied to the general population gave acceptable levels of reliability for parents and adolescents, and low test–retest agreement in the case of children (Boyle et al., 1993). The DICA discriminates between paediatric and psychiatric samples and is moderately related to other measures of psychopathology of children, such as clinical diagnosis (Herjanic & Campbell, 1977). The reliability data from the Spanish adaptation of DICA-R were obtained using a sample of children aged 6–17 years and their parents. Reliability among interviewers scored kappa values of 1.0 in the majority of categories; the test–retest reliability was very good, the majority of the kappa values ranging from good to excellent (Ezpeleta, de la Osa, Doménech, Navarro, & Losilla, in press). The parents were the most stable informants, followed by the children and finally by the adolescents (Ezpeleta, de la Osa, Doménech, Navarro, & Losilla, 1995). There were notable differences depending on the informant and the nature of the information (cognitive vs. observable) (de la Osa, Ezpeleta, Doménech, Navarro, & Losilla, 1996; Ezpeleta et al., 1997, in press).

The degree of functional impairment was measured by the Children's Global Assessment Scale (CGAS; Shaffer et al., 1983), which gives the lowest and the highest level of the psychological adjustment of children and adolescents aged 4–16 years over a specific period of time (in our study, the previous month). The range of values was between 1 (highest level of impairment) and 100 (highest level of adjustment). Scores around 60 or lower indicate an abnormal psychological adjustment (Bird et al., 1990).

Procedure

The children and adolescents were interviewed on two different occasions (test and retest) by different interviewers. Informed written consent was obtained from the parents and verbal consent was obtained from all the children and adolescents included in the study. The interviewers, who had previously been trained in the use of the DICA-R, were unaware of any details relating to the case (they did not have access to the data previously collected by the research team when the interview was carried out at retest, nor did they have access to the assessments carried out by the clinic or hospital staff). The average interval between test and retest was 11 days.

Computer management of the data was carried out using the DAT System 2.0 (Doménech & Losilla, 1995), a relational database management system.

Statistical Analysis

The present study makes use of the concept that Fallon and Schwab-Stone (1994) call *individual reliability*, for a child and for a question. The individual reliability for a child is defined as the subject's ability to provide consistent information in the two interviews (test and retest). The individual reliability for a question is defined as the characteristic property of each item to obtain the same answer when the item is reformulated to the same subject (Fallon & Schwab-Stone, 1994).

According to our hypotheses, the individual reliability of the subject depends on the age, gender, and level of functional impairment, whereas the individual reliability of the question depends on the total number of words of each question, its content, and other specific characteristics of the item concerned.

In order to ascertain the verisimilitude of these hypotheses, we constructed various multiple linear regression models that separately measure the influence of the characteristics of the informants and those of the questions, in the same way as Fallon and Schwab-Stone (1994). These models required an individual reliability measurement as a dependent variable; since this attribute is a function of the number of agreements/disagreements in the test–retest binomial, the kappa statistic (Cohen, 1960) was used. As it is now accepted that this coefficient is dependent on several factors, particularly the degree of skew or the proportion of negative answers for the particular disorder (Kraemer, 1979), the base rate value was included in the regression models in order to correct that bias (Fallon & Schwab-Stone, 1994).

The models that measured the contribution of the characteristics of the subjects included the independent variables of age, gender, and the degree of functional impairment (CGAS), and, as a dependent variable, the kappa value obtained by adding together the answers given by each subject in the two interviews (test and retest).

In the regression model used to study the role played by the attributes of the questions, the independent variables were the total number of words in each question and the different dimensions according to which the questions were classified (depending on whether the given characteristic was present or absent): type of content (internalising vs. externalising), the presence of time concepts, the requirements to specify a group of symptoms, references to the frequency of a symptom, reference to the degree of impairment caused by the disorder, the requirement for the subject to compare himself/herself with his/her peers, and the need to exercise judgement. The individual reliability of each question was calculated by adding together the answers given by all the subjects to each item and then comparing the answers given in the first and the second interview.

Since in practice it has been impossible to define a model to assess how far the characteristics of the children interact with the characteristics of the questions, because it was not possible to obtain a simultaneous measurement of the reliability of the informants and the questions involved, we have opted to specify different regression models. In these models the reliability for each subject was calculated for each type of question; in other words, the reliability for each informant was calculated for each of the characteristics in a given item.

Finally, we have also obtained the weighted sensitivity and specificity of the questions, for each dimension and for each diagnostic category of the DICA-R. For the sensitivity and specificity calculation, time taken on the test was considered as the comparison criterion. We believe that this could be another interesting analysis that was not included in the Fallon and Schwab-Stone (1994) study, since these coefficients are independent of the base rate and easily understandable.

Results

The reliability of the answers given in the DICA-R by the subjects aged 7–17 years is, in the majority of cases, good or excellent. However, there are characteristics of both the informant and the interview which affect this reliability.

Characteristics of the Informant

Table 1 shows the results obtained in the regression model, which reflects the characteristics of the informants as independent variables and the average kappa value of each subject as the dependent variable. The interaction terms were excluded from the final model because they were not significant ($p = .2721$; Doménech, 1996). The base rate, however, was retained as a covariable. The results obtained in estimating this model indicate that the characteristics of the children account for 35% of the variability of the dependent variable ($p < .00005$). The only significant informant characteristic when predicting the kappa values is the level of impairment ($B = 0.004$; $p = .0004$), indicating that the children with a higher level of functional adjustment provide answers that are more stable over time. In the adolescent group, the interactions have also proved to be nonsignificant ($p = .1532$). The characteristics of the adolescent subjects account for 20% of the variability of the dependent variable ($p = .0121$), although none of the characteristics achieved statistical significance.

Characteristics of the Questions

Table 2 shows the average value of the kappa coefficients of the questions, the sensitivity, the specificity,

Table 1
Characteristics of the Subjects that Predict Agreement or Disagreement between Test and Retest

	Final fitted model			
Characteristics	B	SE B	*β*	*p*
Children[a]				
Age	−0.0065	0.0064	−0.1107	.3143
Impairment	0.0045	0.0012	0.4557	.0004
Gender	−0.0134	0.0237	−0.0618	.5763
Base rate	−0.5665	0.2429	−0.2815	.0236
Constant	0.5250	0.1098		
Adolescents[b]				
Age	0.0272	0.0137	0.2498	.0524
Impairment	0.0018	0.0012	0.1963	.1408
Gender	0.0258	0.0354	0.0887	.4696
Base rate	−1.1194	0.3094	−0.4877	.0007
Constant	0.2924	0.2201		

[a] Children: $N = 57$; Contribution of the children's characteristics: $R^2 = .35$ ($p < .00005$); Final fitted model: $R^2 = .41$ ($p < .00005$).

[b] Adolescents: $N = 52$; Contribution of adolescents' characteristics: $R^2 = .21$ ($p < .0121$); Final fitted model: $R^2 = .38$ ($p < .0002$).

Table 2
Average Kappa Coefficient Values for Children and Adolescents in each Diagnostic Category

	Children				Adolescents			
Diagnostic category	Kappa (*SD*)	Sensitivity	Specificity	Base rate	Kappa (*SD*)	Sensitivity	Specificity	Base rate
Demographics	.872 (.116)	.974	.989	—	.908 (.069)	.926	.993	—
Attention deficit/ hyperactivity disorder	.370 (.120)	.502	.929	.191	.510 (.132)	.658	.922	.214
Oppositional/defiant disorder	.397 (.106)	.564	.939	.201	.343 (.130)	.436	.993	.291
Conduct disorder	.607 (.280)	.543	.973	.077	.589 (.256)	.706	.967	.151
Alcohol use and abuse	.934 (.133)	.600	1.000	.003	.801 (.196)	.801	.989	.041
Cigarette smoking	.622 (.415)	.333	.996	.021	.761 (.081)	.796	.976	.185
Glue sniffing	.875 (.307)	—	1.000	.001	.994 (.016)	.000	1.000	.004
Marihuana	.981 (.141)	—	1.000	.001	.956 (.121)	.667	1.000	.006
Street drugs	.970 (.164)	—	1.000	.001	.974 (.118)	.391	1.000	.008
Depressive episode	.465 (.266)	.449	.967	.056	.383 (.164)	.563	.898	.249
Manic episode	.293 (.280)	.237	.971	.037	.354 (.111)	.372	.918	.178
Dysthymic disorder	.103 (.277)	.143	.982	.015	.358 (.109)	.436	.952	.191
Separation anxiety disorder	.333 (.086)	.418	.936	.158	.430 (.139)	.432	.974	.113
Avoidant disorder	.341 (.218)	.267	.965	.059	.482 (.277)	.250	.957	.094
Overanxious disorder	.378 (.103)	.375	.976	.130	.324 (.095)	.486	.917	.325
Phobias	.445 (.177)	.482	.972	.093	.582 (.222)	.325	.964	.059
Obsessive-compulsive disorder	.318 (.237)	.337	.966	.068	.353 (.172)	.387	.945	.104
Post-traumatic stress disorder	.497 (.066)	.364	.993	.048	.217 (.085)	.144	.987	.078
Anorexia nervosa	.717 (.266)	.692	.999	.015	.660 (.142)	.771	.953	.179
Bulimia nervosa	.850 (.287)	.250	.998	.007	.523 (.061)	.755	.962	.086
Enuresis	.376 (.210)	.521	.963	.127	.608 (.251)	.780	.990	.067
Encopresis	.389 (.084)	.524	.962	.062	.701 (.055)	.555	.555	.058
Gender identity (boys)	.859 (.144)	.000	1.000	.002	.508 (.171)	.125	1.000	.016
Gender identity (girls)	.974 (.020)	.625	.995	.020	.547 (.160)	.500	.125	.022
Somatisation	.604 (.400)	.393	.993	.016	.322 (.153)	.329	.974	.056
Psychotic symptoms	.538 (.203)	.448	.985	.044	.362 (.292)	.382	.965	.046

Table 3
Average Kappa Coefficient Values for Children and Adolescents for each Characteristic of the Questions

	Children				Adolescents			
Question dimension	Kappa (*SD*)	Sensitivity	Specificity	Base rate	Kappa (*SD*)	Sensitivity	Specificity	Base rate
Type of content								
Internalising	.531 (.314)	.414	.976	.060	.451 (.264)	.489	.947	.144
Externalising	.687 (.326)	.567	.988	.044	.668 (.271)	.657	.981	.083
Determining time								
No	.654 (.327)	.493	.984	.052	.614 (.295)	.570	.967	.107
Yes	.501 (.315)	.428	.973	.054	.444 (.219)	.485	.966	.139
Symptoms occur in groups								
No	.621 (.330)	.480	.983	.051	.578 (.289)	.557	.967	.109
Yes	.547 (.353)	.567	.956	.100	.547 (.318)	.571	.947	.269
Frequency of symptoms								
No	.608 (.329)	.477	.982	.054	.562 (.287)	.549	.965	.114
Yes	.737 (.317)	.623	.989	.024	.731 (.266)	.736	.989	.066
Intensity-impairment of disorder								
No	.623 (.330)	.495	.994	.054	.591 (.289)	.573	.965	.115
Yes	.593 (.334)	.317	.985	.030	.454 (.256)	.393	.974	.081
Comparison with peers								
No	.629 (.329)	.486	.983	.051	.579 (.291)	.567	.967	.113
Yes	.331 (.240)	.385	.973	.055	.498 (.212)	.375	.963	.077
Child's judgement								
No	.668 (.328)	.834	.988	.012	.627 (.292)	.679	.979	.090
Yes	.522 (.313)	.432	.975	.099	.469 (.251)	.473	.953	.132

Table 4
Characteristics of Questions that Predict Agreement or Disagreement between Test and Retest

	Final model			
Characteristics	B	SE B	β	p
Children[a]				
Length	−0.0028	0.0014	−0.0701	.0479
Content	0.0838	0.0226	0.1258	.0002
Time	−0.1641	0.0268	−0.2058	.0000
Clustering	−0.0551	0.0913	−0.0198	.5465
Frequency	−0.0223	0.0407	−0.0189	.5835
Impairment	−0.0532	0.0407	−0.0506	.1915
Comparison	−0.1960	0.0660	−0.1012	.0031
Judgement	−0.0913	0.0277	−0.1303	.0010
Base rate	−1.6021	0.1381	−0.3865	.0000
Constant	0.7700	.0297		
Adolescents[b]				
Length	−0.0007	0.0012	−0.0194	.5796
Content	0.1663	0.0199	0.2839	.0000
Time	−0.1247	0.0239	−0.1774	.0000
Clustering	−0.0321	0.0818	−0.0129	.6952
Frequency	0.0388	0.0354	0.0371	.2744
Impairment	−0.1158	0.0360	−0.1238	.0014
Comparison	0.0119	0.0588	0.0069	.8399
Judgement	−0.0724	0.0246	−0.1168	.0034
Base rate	−0.4429	0.7290	−0.2050	.0000
Constant	0.6002	.0263		

[a] Children: $N = 57$; Contribution of the question characteristics: $R^2 = .14$ ($p < .00005$); Final fitted model: $R^2 = .28$ ($p < .00005$).

[b] Adolescents: $N = 52$; Contribution of the question characteristics: $R^2 = .22$ ($p < .00005$); Final fitted model: $R^2 = .26$ ($p < .00005$).

and the base rate, calculated for each diagnostic category in the interview. We have calculated the kappa statistics from the average value of the kappa coefficients for each question. As expected, the questions on demographic data obtained answers of outstanding reliability. Similarly, for both types of informants the average kappa coefficient values were very high in the case of disorders relating to the use and abuse of drugs and alcohol; in the case of children, eating disorders and gender identity also gave very high levels of agreement. In this group, disorders due to attention deficit/hyperactivity disorder (ADHD), manic episode, dysthymic disorder, anxiety disorders, and elimination disorders gave the lowest kappa values. The sensitivity values were generally low, except for bulimia nervosa and gender identity (boys), which gave lower values than expected by kappa. Specificity values were very high in all cases. The basic rate values were less than .10, except for ADHD, oppositional/defiant disorder, separation anxiety disorder, overanxious disorder, and enuresis. In adolescents, the poorest results of agreement were obtained in cases of oppositional/defiant disorder, mood disorders, overanxious disorder, obsessive-compulsive disorder, post-traumatic stress disorder, somatisation, and psychotic symptoms. Similar results were found for sensitivity, except for glue sniffing, street drugs, and gender identity (boys). These sensitivity values were, in the majority of cases, poor. The base rate values were higher than in children, except for enuresis. Specificity values were also very high in all the categories.

Table 3 shows the average kappa coefficient, the sensitivity, the specificity, and the base rate values calculated for each dimension of the questions. The results indicate that for both types of informants, the questions that obtain the most reliable answers (kappa) are those that have an externalising content, include no time concepts, do not work out clustering of symptoms, involve determining the frequency of conduct occurring, do not require the degree of impairment of the disorder or the comparison with peers to be determined, and do not ask the informant to exercise his/her judgement. The poorest kappa value was obtained when the question required the younger children to compare their conduct or feelings with those of their peer group. The sensitivity values were low in general, and higher in adolescents than in children. In all cases, specificity values were higher than .90.

Table 4 shows the results obtained in the final regression model, which reflects the question characteristics as independent variables and the kappa coefficient of each question as the dependent variable. The data at the top of the table correspond to the children in the sample and indicate that the question characteristics account for 14% ($p < .00005$) of the variability of the dependent variable. These results suggest that the length of the question, content (internalising is more difficult), time concepts, the requirement for the subject to compare him/herself with others and to exercise judgement are the significant characteristics in the model. For adolescents, the results indicate that the question characteristics account for 22% ($p < .00005$) of the variability. The significant dimensions in this model were: internalising content, the presence of time concepts, determination of the degree of impairment caused by the disorders, and the need for the subject to exercise judgement.

Models of Interaction between the Characteristics of the Children and Those of the Questions

The results indicate that none of the characteristics of the children or the adolescents is relevant in explaining the reliability of answers according to the various characteristics of the questions, except the level of functional impairment of the younger children (7–12 years) in the regression model defined for the questions containing externalising attributes ($B = 0.0025$; $p = .0332$). The presence of functional impairment in these children leads to lower reliability in the case of these questions.

Discussion

Children and adolescents are reliable informants when answering the DICA-R questions, but as we had expected, there are certain characteristics of young informants and of the interview itself that affect the stability of answers during the two interviews (test–retest).

With regard to the question characteristics, a first hypothesis supposes that exposing the younger subjects to such complex questions as those which constitute a diagnostic interview could be beyond their cognitive abilities. For example, according to Piaget's theory (Piaget & Inhelder, 1984), subjects acquire the concept of time during the stage of concrete operational thought. This means that subjects over the age of 7 should be able to understand and deal with the basic notions of time, as well as the commonly used units of physical measurement. However, the data obtained in the present study, as well as in others carried out using the DISC (Breton et al., 1995; Fallon & Schwab-Stone, 1994), show that informants aged 7–17 years find it difficult to handle questions that include time concepts. In order to account for this apparent contradiction, we should bear in mind that Piaget worked with time calculations in very concrete experiments, and did not study the notion of time on a more abstract level. However, during assessment with diagnostic interviews, the informant is forced to make a mental representation of a set of words and concepts without the help of any concrete reality or experience, that is to say, he/she is forced to perform a cognitive task that often exceeds his/her capacity for thought. Another important factor is that many of the items included in these protocols force the subject to think in retrospective sequences, which is more complex both for children and adolescents, even when they have the support of temporal guidelines. Similarly, many questions referring to the duration of a particular conduct are also based on time intervals (days, weeks, or months) that vary from question to question, or include vague and complex notions (such as "sometimes" or "for a long time") that pose difficulties regarding the subjects' level of comprehension (Breton et al., 1995; Valla, Bergeron, Bérubé, Gaudet, & St Georges, 1994).

Regarding the results concerning the length of the questions, some contemporary cognitive theories suggest that the process of encoding and storing information becomes more complex and efficient as the subject develops, which would predict a variation in the reliability of the answers of the children depending on the length of the questions used (this is particularly likely to be the case if we bear in mind that longer questions are usually those that include more than one concept or idea). Indeed, we have observed that the younger children are those who have had greatest difficulty with the long items; however, the reliability of answers in the adolescent subjects was not found to depend on this characteristic.

As for the type of content in the questions, results indicate that, in general, questions with an externalising content, which do not require subjects to compare their behaviour with that of their peers and do not require them to exercise their judgement, show greater reliability in the answers. These characteristics reduce the level of complexity of the questions and the informants are more reliable in their answers because they are not forced to be highly introspective or to exercise judgement concerning their own behaviour and experiences. Accordingly, several studies have shown that there is greater agreement, both in the case of parents and children, when the question deals with observable as opposed to inner elements (Ezpeleta et al., 1995; Silverman & Eisen, 1992).

Similarly, questions which require the subject to indicate the frequency of symptoms have obtained the most reliable answers. It should be borne in mind that in the case of these questions, the informant can choose from several alternative answers that reflect the various frequency intervals during which a certain type of behaviour occurred (for example: once, twice, 3–4 times, 5–9 times, more than 10 times), which may help the subject to be clear about the alternative that best represents his/her conduct.

It is interesting to note that both the children and the adolescents obtained the lowest reliability scores in the same sections of the interview (Dysthymic disorder, manic episode, overanxious disorder, and obsessive-compulsive disorder). These are precisely the sections that account for the greatest number of internalising questions and that require the exercise of judgement on the part of the subject, a fact that could explain the difficulties experienced by the informants in the case of these questions and the resulting drop in reliability. On the other hand, both children and adolescents gave very reliable answers in the case of drug and alcohol use-abuse disorders, eating disorders, and those relating to gender identity. These results would reflect a good level of reliability for the answer "no", since these disorders were not prevalent in the sample involved in the present study.

The results concerning subject characteristics are also worthy of comment. As regards gender, our results coincide with those of other studies that have been carried out to date (Canino et al., 1987; Edelbrock et al., 1986; Klein, 1991; Rapee et al., 1994; Reich et al., 1982): there were no differences in the reliability of answers between boys and girls aged 7–17. Unlike the results obtained by other authors (Edelbrock et al., 1985, 1986; Schwab-Stone et al., 1994; Silverman & Eisen, 1992), our study found no differences relating to the age of the informants, probably due to the fact that the age range of the children included in each of the regression models was too small to obtain such differences. In general, we found that the children aged 7–12 years were more consistent in their answers than the adolescent subjects aged 13–17 years, as was previously reported (Ezpeleta et al., in press).

The level of functional impairment in the younger subjects, however, was a variable that explained the degree of reliability in the answers given by the younger informants. Specifically, our results suggest that the children with the highest level of impairment are those who give the least reliable information. We know that those children with the greatest impairment are likely to have the most severe or the greatest number of symptoms. The method that we applied for the test–retest reliability study is one of the most stringent, since in the period between the test and the retest many changes may have occurred (including a decrease or an increase in the symptoms, a change in the symptoms, the subject having learned to say no in order to make more rapid progress in the interview, changes in the interpretation of the questions due to the fact that in the retest the complete structure of the interview or the aim of the study is known, etc.) (Robins, 1985). These children who show the greatest impairment may, for that reason, be more susceptible to factors that may occur between test and retest.

Finally, we have also observed that the children with functional impairment are less reliable in the case of questions with externalising content. In fact, Verhulst, Eussen, Berden, Sanders-Woudstra, and Van der Ende (1993), in a study that monitored children of the general population over a period of 6 years, found that the majority of subjects whose level of general psychological disorder persisted over time were diagnosed as having externalising disorders and showed aggressive or antisocial behaviour; on the contrary, children with internalising disorders tended to improve over time. It seems, therefore, that externalising disorders, because of their usually long duration, might lead to serious overall impairment. These children (with ADHD, oppositional/defiant disorder, or conduct disorder), who have suffered from their condition over a long period, particularly if they are young, would be expected to give the least reliable information. Moreover, the characteristics inherent in these disorders (lying, lack of concentration, distractedness, impulsiveness) should not be overlooked, since they might also have a bearing on the subjects' giving of information.

We also wish to indicate some methodological implications for the results obtained in this work. Perhaps the most difficult problem in using the kappa reliability coefficient is that its value varied with some parameters, such as the illness base rate. Some researchers argue that in clinical studies one can compare kappa values for those disorders with base rates in the middle range without as much concern, because in these cases the problem of the base rate dependence of kappa is relatively small (Spitznagel & Helzer, 1985). However, in population studies in which base rates are typically low for some disorders, the problem becomes more serious.

We have not been able to find a single measure that is completely independent of the base rate for the great variety of conditions under which agreement can be measured. For this reason, in our study, we have used kappa coefficients, and attempts were made in the regression models to account for the base rate. However, other parameters that could also affect the kappa values were not accounted for (e.g. there is no standardised way to account for the asymmetry of the N × N table [Feinstein & Cicchetti, 1990]).

In the descriptive section we have also used two separate indices, sensitivity and specificity, because we consider that this could be a unique way to resolve the omnibus kappa problem. Usually, sensitivity is defined as the fraction of time a test will take to make a positive diagnosis when a disorder is present, and specificity is defined as the fraction of time a test will take to make a negative diagnosis when the disorder is really absent. In this paper we have calculated sensitivity and specificity for each dimension and for each diagnostic category of the questions of the DICA-R, considering time taken on test as the comparison criterion. We believe these individual values of sensitivity and specificity make important contributions when results are interpreted for studies of test–retest variability. These coefficients will indicate the consistency of the two interviews (test and retest) when they go in the opposite directions of positive and negative decisions. The distinction can help the reader to decide about the persuasiveness of the individual results, and will also help researchers to design further work to decrease the test–retest disparities in positive, negative, or both directions (Cicchetti & Feinstein, 1990).

In summary, since no single index will be satisfactory for the purposes of understanding what happens in the test–retest process or improving the agreement, we think that the kappa values should always be accompanied by separate individual values of sensitivity and specificity.

On the other hand, we must underline that the interactions between the questions' attributes and the children's characteristics cannot be tested in the present work because of the methodological limitations. For example, there may be significant interactions between age and the various question attributes, but it has been impossible to define an appropriate model to assess these parameters.

Finally, we would also like to suggest the existence of alternative and powerful methods of evaluating the statistical and practical significance of the hypothesis confronted in this paper, such as the extent to which the characteristics of the subjects (age, gender, and degree of functional impairment) influence test–retest reliability. For example, we know that techniques developed in the context of structural equation modelling are useful methods for detecting and describing population heterogeneity that cannot be handled in regular multiple-group analysis. In this way, Muthén (1989) provides an interesting overview of the methodology than can address population heterogeneity on test structures. Although real-world applications still require the development of more tailored modelling that takes into account the special features of some questions in the research, it could be interesting to carry out further works examining the extent to which model parameters vary across different sample strata using these techniques.

On the whole, the results obtained in this study indicate that children can provide consistent information needed for their psychological assessment when they are asked directly. Nevertheless, the cognitive level of the child is a variable that must always be taken into account when designing and using structured and semistructured protocols. It is true that both the interviews and the classifications on which they are based share the criticism concerning low sensitivity to the developmental level of the subject, since they deal with symptoms in a static fashion and fail to take into account developmental changes and the processes of adjustment, organisation, and interaction between those changes (Beitchman, Wekerle, & Hood, 1987; Cantwell & Baker, 1989; Dalton, Forman, Daul, & Bolding, 1987; Ezpeleta, 1995; Ezpeleta et al., in press). We have verified that before the age of 12 years, children give consistent information in answer to many of the DICA-R questions; we have also observed, however, that at this age some children (and also older subjects) may find it difficult to understand all the questions in this protocol. Similarly, the degree of functional impairment of the younger children should be taken into account when gathering information. A consideration of these variables is a necessary prerequisite and a guideline for the development of new interview protocols.

Acknowledgements—This work was carried out with the assistance of the Spanish Ministry of Education and Science DGICYT PM91-0209 and DGICYT PM95-126.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd edn. revised). Washington, DC: Author.

Beitchman, J. H., Wekerle, C., & Hood, J. (1987). Diagnostic continuity from preschool to middle childhood. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 694–699.

Bird, H. R., Yager, T. J., Staghezza, B., Gould, M. S., Canino, G., & Rubio-Stipec, M. (1990). Impairment in the epidemiological measurement of childhood psychopathology in the community. *Journal of the American Academy of Child and Adolescent Psychiatry*, *29*, 796–803.

Boyle, M. H., Offord, D. R., Racine, Y., Sandorf, M., Szatmari, P., Fleming, J. E., & Price-Munn, N. (1993). Evaluation of the Diagnostic Interview for Children and Adolescents for use in general population samples. *Journal of Abnormal Child Psychology*, *21*, 663–681.

Breton, J., Bergeron, L., Valla, J., Lépine, S., Houde, L., & Gaudet, N. (1995). Do children aged 9 through 11 years understand the DISC version. 2.25 questions? *Journal of the American Academy of Child and Adolescent Psychiatry*, *34*, 946–954.

Canino, G. J., Bird, H. R., Rubio-Stipec, M., Woodbury, M. N., Ribera, J. C., Huertas, S. E., & Sesman, M. J. (1987). Reliability of child diagnosis in a Hispanic sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 560–565.

Cantwell, D. P., & Baker, L. (1989). Stability and natural history of DSM-III childhood diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*, 691–700.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: I. Resolving the paradoxes. *Journal of Clinic Epidemiology*, *43*, 551–558.

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

Dalton, R., Forman, M. A., Daul, G. C., & Bolding, D. (1987). Psychiatric hospitalization of preschool children: Admission factors and discharge implications. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 308–312.

de la Osa, N., Ezpeleta, L., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (1996). Fiabilidad entre entravistadores de la DICA-R. *Psicothema*, *8*, 359–368.

Doménech, J. M. (1996). *Análisis multivariante: Modelos de regresión*. Barcelona, Spain: Signo.

Doménech, J. M., & Lossila, J. B. (1995). *Sistema DAT: Gestor de datos cientificos. Manual de referencia*. Campus de Bellaterra, Barcelona: Laboratori d'Estadística Aplicada i de Modelització, Universitat Autònoma de Barcelona.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Conover, N. C., & Kalas, R. (1986). Parent–child agreement on child psychiatric symptoms assessed via structured interview. *Journal of Child Psychology and Psychiatry*, *27*, 181–190.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the Psychiatric Interview of the Child. *Child Development*, *56*, 265–275.

Ezpeleta, L. (1995). Las entravistas estructuradas en el diagnóstico psicopatológico infantil. In J. Rodríguez (Ed.), *Psicopatología del niño and del adolescente* (pp. 305–329). Sevilla, Spain: Servicio de Publicaciones de la Universidad de Sevilla.

Ezpelata, L. (1996). Entrevistas estructuradas para el diagnóstico psicopatológico en niños and adolescents. In J. M. G. Alberca & C. G. Prieto (Eds.), *Manual práctico de psicología clínica and de la salud* (pp. 95–120). Sevilla, Spain: Publicaciones del Dentro Clínico los Naranjos.

Ezpeleta, L., de la Osa, N., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (1995). La Diagnostic Interview for Children and Adolescents-Revisada (DICA-R): Acuerdo diagnóstico entre niños/adolescentes y sus padres. *Revista de Psiquiatría de la Facultad de Medicina de Barcelona*, *22*, 153–163.

Ezpeleta, L., de la Osa, N., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (in press). Test–retest reliability of the Diagnostic Interview for Children and Adolescents (DICA-R). *Psicothema*.

Ezpeleta, L., de la Osa, N., Judez, J., Doménech, J. M., Navarro, J. B., & Losilla, J. M. (1997). Diagnostic agreement between clinicians and the Diagnostic Interview for Children and Adolescents—DICA-R—in an outpatient sample. *Journal of Child Psychology and Psychiatry*, *38*, 431–440.

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

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinic Epidemiology*, *43*, 543–549.

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., Herjanic, M., Brown, F., & Wheatt, T. (1975). Are children reliable reporters? *Journal of Abnormal Child Psychology*, *3*, 41–48.

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

Kashani, J. H., Orvaschel, H., Burk, J. P., & Reid, J. C. (1985). Informant variance: The issue of parent–child disagreement. *Journal of the American Academy of Child and Adolescent Psychiatry*, *24*, 437–441.

Klein, R. G. (1991). Parent–child agreement in clinical assessment of anxiety and other psychopathology: A review. *Journal of Anxiety Disorders*, *5*, 187–198.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, *44*, 461–472.

La Greca, A. M., & Stone, W. L. (1992). Assessing children through interviews and behavioral observations. In C. E. Walker & M. C. Roberts (Eds.), *Handbook of clinical child psychology* (2nd edn.) (pp. 63–83). New York: Wiley Interscience.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology*, *19*, 136–143.

Loeber, R., Green, S. M., Lahey, B. B., & Stouthamer-Loeber, M. (1991). Differences and similarities between children, mothers, and teachers as informants on disruptive child behavior. *Journal of Abnormal Child Psychology*, *19*, 75–95.

Muthén, B. (1989). Latent variable modeling in heterogenous populations. *Psychometrika*, *54*, 557–585.

Piaget, J., & Inhelder, B. (1984). *Psicología del niño* (L. Hernández Alfonso, Trans.). Madrid: Morata. (Original work published 1969.)

Rapee, R. M., Barret, P. M., Dadds, M. R., & Evans, L. (1994). Reliability of the DSM-III-R childhood anxiety disorders using structured interview: Inter-rater and parent–child agreement. *Journal of the American Academy of Child and Adolescent Psychiatry*, *33*, 984–992.

Reich, W., Herjanic, B., Welner, Z., & Gandhy, P. R. (1982). Development of a structured psychiatric interview for children: Agreement on diagnosis comparing child and parent interviews. *Journal of Abnormal Child Psychology*, *10*, 325–336.

Reich, W., Shayka, J., & Taibleson, Ch. (1991). *Diagnostic Interview for Children and Adolescents-Revised version 7.2* (L. Ezpeleta, Trans.). Unpublished manuscript, Washington University, Division of Child Psychiatry, St. Louis.

Robins, L. N. (1985). Epidemiology: Reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry*, *42*, 918–924.

Rutter, M., & Graham, P. (1968). The reliability and validity of the psychiatric assessment of the child. I. Interview with the child. *British Journal of Psychiatry*, *114*, 563–579.

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. *American Journal of Psychiatry*, *151*, 1048–1054.

Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., & Aluwahlia, S. (1983). A Children's Global Assessment Scale (CGAS). *Archives of General Psychiatry*, *40*, 1228–1231.

Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Piacentini, J., Davies, M., Conners, K., & Regier, D. (1993). The Diagnostic Interview Schedule for Children-Revised version (DISC-R): I. Preparation, field testing, inter-rater reliability and acceptability. *Journal of the American Academy of Child and Adolescent Psychiatry*, *32*, 643–650.

Silverman, W. K., & Eisen, A. R. (1992). Age differences in the reliability of parent and child reports. Child anxious symptomatology using a structured interview. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 117–124.

Silverman, W. K., & Kearney, C. A. (1992). Listening to our clinical partners: Informing researchers about children's fears and phobias. *Journal of Behaviour Therapy and Experimental Psychiatry*, *23*, 71–76.

Spitznagel, E. D., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, *42*, 725–728.

Sylvester, C. E., Hyde, T. S., & Reichler, R. J. (1987). The

Diagnostic Interview for Children and the Personality Inventory for Children in studies of children at risk for anxiety disorders or depression. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 668–686.

Valla, J. P., Bergeron, L., Bérubé, H., Gaudet, N., & St Georges, M. (1994). A structured pictorial questionnaire to assess DSM-III-R based diagnoses in children (6–11 years): Development, validity and reliability. *Journal of Abnormal Child Psychology*, *22*, 403–423.

Verhulst, F. C., Althaus, M., & Berden, G. F. M. G. (1987). The Child Assessment Schedule: Parent–child agreement and validity measures. *Journal of Child Psychology and Psychiatry*, *28*, 455–466.

Verhulst, F. C., Eussen, M. L., Berden, G. F., Sanders-Woudstra, J., & Van der Ende, J. (1993). Pathways of problem behaviours from childhood to adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry*, *32*, 388–396.

Weissman, M. M., Wickramaratne, P., Warner, V., John, K., Prusoff, B. A., Merikangas, K. R., & Gammon, G. D. (1987). Assessing psychiatric disorders in children. *Archives of General Psychiatry*, *44*, 747–753.

Welner, Z., Reich, W., Herjanic, B., Jung, K. G., & Amado, H. (1987). Reliability, validity and parent–child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 649–653.

Werry, J. S. (1992). Child psychiatric disorders: Are they classifiable? *British Journal of Psychiatry*, *161*, 472–480.

Williams, S., McGee, R., Anderson, J., & Silva, P. A. (1989). The structure and correlates of self-reported symptoms in 11-year-old children. *Journal of Abnormal Child Psychology*, *17*, 55–71.

Young, J. G., O'Brien, J. D., Gutterman, E. M., & Cohen, P. (1987a). Structured diagnostic interviews for children and adolescents. Introduction. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 611–612.

Young, J. G., O'Brien, J. D., Gutterman, E. M., & Cohen, P. (1987b). Research on the clinical interview. *Journal of the American Academy of Child and Adolescent Psychiatry*, *26*, 613–620.

Zahner, G. E. P. (1991). The feasibility of conducting structured diagnostic interviews with preadolescents: A community field trial of the DISC. *Journal of the American Academy of Child and Adolescent Psychiatry*, *30*, 659–668.

Manuscript accepted 8 January 1998