British Journal of Psychiatry (1991), **159** (suppl. 11), 36–40

Reliability and Validity of Two Self-Rating Scales in the Assessment of Childhood Depression

T. FUNDUDIS, T. P. BERNEY, I. KOLVIN, O. O. FAMUYIWA, L. BARRETT, S. BHATE and S. P. TYRER

A comparison was made of the reliability and validity of two self-rating scales, the Children's Depression Inventory (CDI) and Depression Self-Rating Scale (DSRS), in the diagnosis of depression in 93 children (aged 8–16 years) attending a university child psychiatry department. The two scales were of comparable merit but had only moderate discrimination between depressed and non-depressed children, with each scale having a misclassification rate of 25%. Better agreement was obtained in more verbally intelligent children, irrespective of age. Girls scored higher on the instruments than boys. No significant relationship was found between teacher assessment of classroom behaviour and the two self-rating depression instruments.

Childhood depression has attracted mounting attention over the past decade (Schulterbrandt & Raskin, 1977; Kovacs, 1981; Kazdin & Petti, 1982; Puig-Antich & Gittelman, 1982; Kolvin *et al*, 1984). One important advance has been the development of self-rating instruments designed to screen for affective disorder in childhood (Kovacs & Beck, 1977; Birleson, 1981; Tisher & Lang, 1983). The Children's Depression Inventory (CDI) has been extensively researched in the US (Kovacs & Beck, 1977; Kovacs, 1981; Saylor *et al*, 1984; Helsel & Matson, 1984), but there has been no comparable work in the UK. Although Birleson (1981) developed the Depression Self-Rating Scale (DSRS) in Edinburgh almost a decade ago, there has been only one reported study on its further use in a British population (Firth & Chaplin, 1987). The present study assesses the reliability and validity of the CDI and DSRS as screening measures for affective disorder in childhood and adolescence.

The aims of the research were to ascertain:

(a) the agreement between the CDI and the DSRS
(b) their validity in relation to clinical assessment of depression in children referred to a child psychiatry clinic
(c) the relationship between child self-ratings of depression and teacher ratings of classroom functioning.

Method

The CDI is a 27-item self-report inventory standardised for children between eight and 13 years of age (Kovacs & Beck, 1977; Kovacs, 1981). Each item includes three descriptions of increasing severity, for example, "I am sometimes sad", "I am often sad", "I am always sad". The child is asked to indicate which statement most closely applies to him/her with regard to the past two weeks. Responses are scored on a scale from nought to two where two represents the severe form of the particular symptom. The CDI is amenable to short-form administration (Carlson & Cantwell, 1980*b*) and this method was used in the present study. Some of the wording was modified for use with British children.

The DSRS is an 18-item inventory developed on a Scottish population. Each item comprises a statement and the child is asked to indicate whether this applies to him/her "most of the time", "sometimes" or "never". "Sometimes" scores one, "most" or "never" score nought or two, depending on the positive or negative polarity of the item.

The study was conducted at the Nuffield Department of Child and Adolescent Psychiatry and is part of a larger study – the Newcastle Child Depression Project, the details of which are described in the first paper of this volume (Kolvin *et al*, 1991). A consecutive series of referred children aged between eight and 16 years were asked by their psychiatrist to complete a shortened form of the CDI, comprising 14 items, at their first appointment. The psychiatrist was instructed to ensure that the child was able to read the statements. Where the child's reading ability was inadequate for the task the psychiatrist read each statement to the child who was then asked to tick the statements which he/she thought most applied. Nearly all subjects proved adequate for the task.

On the basis of their scores on the CDI, subsamples were selected for inclusion in the study; within four weeks these subsamples were given the DSRS and the remaining half of the CDI. They were also given the Crichton/Mill Hill Vocabulary Scale as a simple measure of intelligence (Raven *et al*, 1976).

On this latter occasion, the child had a psychiatric interview carried out independently by a senior psychiatrist who had no knowledge of the results of the screen instruments. A description of the standard psychiatric interviews and their reliability and validity are described in an accompanying paper by Kolvin *et al* (1991).

The sample

Three hundred and sixteen referred children were 'screened' but this sample was reduced by 41 as follows. In 22 cases

(7%) the families did not attend the follow-up interview; six children (2%) were already on medication (for epilepsy or diabetes) and in two cases (0.6%) parental data were not available as the children were in care; a further 11 cases (3.5%) declined to be included in a parallel study which involved the dexamethasone suppression test – a test for disturbed endocrine function (Tyrer *et al*, 1991, this volume). This left 275 (90%) of the original 316 cases.

Psychiatric assessments could not be carried out on all 275 cases, so roughly equal numbers of high scorers and low scorers on the CDI were selected for interview. A score of less than nine on the short form of the CDI constituted a low score, a score of nine or more a high score. This cut-off score was based on a pilot study of the CDI on 79 cases comprising 66 residing in hospital units and 13 in a social services residential assessment unit. The mean score obtained on this group was 6.9 with a standard deviation of 4.0. The cut-off score of nine was half a standard deviation above the mean. This cut-off score is similar to that used by Carlson & Cantwell (1980*b*) who found that a cut-off score of eight or more proved a satisfactory criterion for screening children with depressive disorder. In our main study half (49 of 100) of the high scorers and one-quarter (44 of 175) of the low scorers were selected randomly. These 93 cases were studied intensively.

Teacher assessment

Teachers were asked to complete two questionnaires on each of the children included for intensive study:
(a) the Conners Teacher Questionnaire – a 39-item checklist which assesses four factors; conduct, inattentive-passivity, tension-anxiety, and hyperactivity (Conners, 1973)
(b) the Newcastle Educational Questionnaire – a simple five-point rating scale on three areas of classroom attainment; reading/language, number/mathematics and attitude to school work (Fundudis *et al*, 1979). To check reliability the results on these dimensions were analysed by product–moment correlation coefficients. This simple statistic was supplemented by more sophisticated analyses of variance and intraclass correlation coefficients, which estimate more accurately the value of true and error variance. Furthermore, the intraclass correlation gives essentially an average correlation across assessors. Inter-rater reliability was based on ratings from two teachers on a random sample of 20 ordinary junior school children. The product–moment correlation coefficients were 0.62 on reading/language attainment, 0.65 on number/mathematics attainment and 0.61 on attitude to school work. Analysis of variance gave highly significant differences between children but none between raters. Intraclass correlation coefficients were 0.61, 0.65 and 0.69 for reading/language, number/mathematics attainment and attitude to school work respectively. Test–retest reliability was determined on the same sample of children over a three-month interval; this yielded product–moment correlation coefficients of 0.84 on reading/language attainment; 0.76 on number/mathematics attainment and 0.85 on attitude to school work. Analysis of variance again gave highly significant differences between the ratees but not between the raters. Intraclass correlation for reading/language attainment, number/mathematics attainment and attitude to school were 0.84, 0.69 and 0.85 respectively. In the case of secondary school children, the likelihood is that teachers know the children less well than in junior schools and that the reliability of teacher ratings might be lower.

Criterion validity was checked by correlating the teacher ratings on reading/language attainment and children's scores on a standardised reading test (Scottish Council for Research in Education, 1967) administered to the main sample of 93 children by an independent interviewer. The product–moment correlation coefficient was found to be 0.68.

Results

Of the 93 children (46 boys, 47 girls) selected for in-depth assessment, 70 were out-patients and 23 were day-patients or in-patients. Their mean age was 12.5 years and their mean vocabulary quotient was 92.8. This sample can be regarded as representative of the patients attending the child–adolescent service in Newcastle.

Reliability and validity

The two screening instruments – the CDI and DSRS – were evaluated for their internal consistency. Split-half reliability was found to be 0.88 on the CDI (Cronbach's alpha) which compared favourably with Kovacs' finding of 0.86 (Kovacs, 1981) and with other studies (Saylor *et al*, 1984). On the DSRS a split-half reliability of 0.82 was obtained which is consistent with the 0.86 reported by Birleson (1981). The estimates of reliability for the full tests using the Spearman–Brown Formula (Guilford, 1954) were 0.92 and 0.90 respectively.

The CDI was also examined for its stability by administering the two parallel halves of the instrument to the same group of 93 children with a three to four week interval between the two halves. This gave a reliability coefficient of 0.71 which is similar to the test-retest reliability found by Kovacs (1981) for a similar time interval, based on the administration of the full instrument. In addition to demonstrating reliability, these findings suggest that the mood state of the group fluctuated only moderately over that brief span of time which coincided with the initial phase of referral. This is consistent with the findings of other studies, (Kovacs *et al*, 1984; Kovacs, 1985; Testiny & Lefkowitz, 1982). We did not examine for stability of depressive symptoms at a subsequent phase of the condition when greater degrees of fluctuation might be expected to occur (Berney *et al*, 1981).

The independent external criterion for assessing validity was the clinical diagnostic rating for the presence and severity of depression as outlined and defined in the Standardised Psychiatric Interview (Goldberg *et al*, 1970). These issues are fully discussed in a preceding paper (Kolvin *et al*, 1991). The interview was conducted by a senior child psychiatrist who was 'blind' to the screening scores. Kappa coefficients (Cohen, 1960) were used for assessment, and moderate correspondence was found between psychiatric diagnosis and categorisation based on the CDI

(kappa = 0.46; maximum kappa 0.85); a similar finding was obtained for the DSRS (kappa 0.42; maximum kappa 0.89). (Maximum kappa is the maximum value of agreement which can be attained in terms of the number of categories and the number of times each rater or judge chooses the category (Cohen, 1960).) Although these kappa coefficients are significant ($P<0.01$) they represent only a moderate concurrent validity. The product–moment correlation coefficient between the CDI and DSRS measures was 0.75.

Teacher assessment and self-rating measures for depression

Ratings by teachers are an invaluable source of information about child dysfunction; teachers' classroom observations were therefore investigated to see if they could provide indicators of childhood depression. No significant agreement was found between the self-rating measures of depression (CDI and DSRS) and the teacher assessments of classroom behaviour (The Conners Scale and the Newcastle Educational Questionnaire). However, children diagnosed as depressed by the clinicians were found to have significantly higher hyperactivity ($P<0.02$) and conduct disturbance scores ($P<0.02$) on the Conners Scale than the non-depressed group.

Utility of CDI and DSRS

A way of expressing the validity of screening instruments derives from epidemiology (Diamond & Lilienfeld, 1962; Goldberg, 1972). The extent to which a screening instrument succeeds (or fails) in the accurate identification of true cases in relation to an external criterion can be measured in terms of three related indices: sensitivity, specificity and misclassification.

Sensitivity is the capacity of a screening instrument to select those who have the disorder. It is calculated by expressing the number of subjects reaching the criteria for the disorder concerned using the instrument as a percentage of all the sample who truly have the disorder.

Specificity is the capacity of a screening instrument to exclude those who do not have the disorder. It is calculated by determining the number of subjects reaching the criteria concerned who are subsequently found not to have the disorder.

Misclassification is calculated by expressing the number of false negatives and false positives (i.e. those wrongly labelled) as a percentage of the total number of subjects studied.

In order to obtain a more adequate representation of the efficiency of the two screening instruments, the rates on the above three indices were not limited to the sub-sample of 93 randomly selected children who were interviewed in depth. The rates were determined by recalculating back to the original sample of 275 children who completed the CDI and DSRS. It has been pointed out that it may be necessary to use a stratified sampling strategy if the prevalence of the disorder chosen is below 40%, to avoid unnecessary work (Goldberg, 1986). If this is done, however, it is essential to weight the data back to the original sample of consecutive subjects, otherwise the estimate of specificity will be too low and the estimate of sensitivity will be too high.

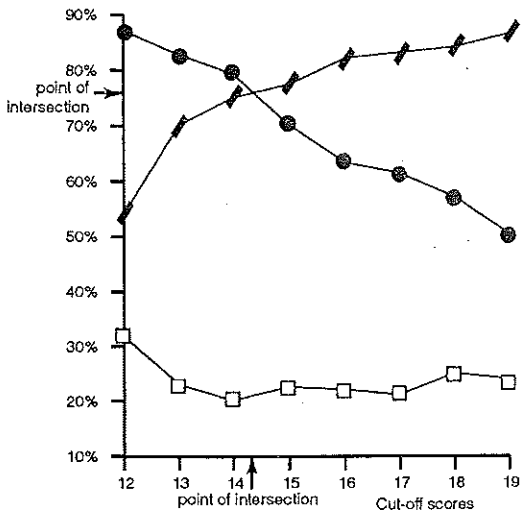

Fig. 1 Percentage rates of sensitivity (●), specificity (▮) and misclassification (□) in relation to different cut-off socres on CDI.

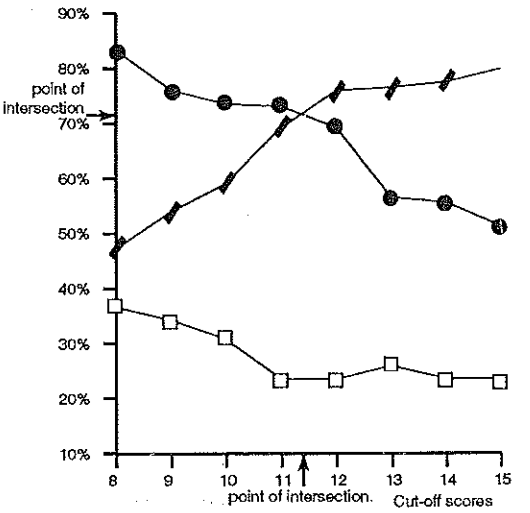

Fig. 2 Percentage rates of sensitivity (●), specificity (▮) and misclassification (□) in relation to different cut-off scores on DSRS.

Figures 1 and 2 show how these three indices vary in relation to different potential cut-off scores on the full version (27-items) of the CDI and the DSRS respectively. The graphical presentation is based on that used by Mari & Williams (1985). In addition, it includes a misclassification curve. The point at which the graphs for sensitivity and specificity intersect is a simple way of

Table 1
Cut-off scores for optimum discrimination on CDI and DSRS as determined by point of intersection for sensitivity and specificity and misclassification rate

	CDI			DSRS		
	Cut-off score	Intersection of sensitivity and specificity %	Misclassification rate %	Cut-off score	Intersection of sensitivity and specificity %	Misclassification rate %
Whole group	15	77	23	11	75	25
Sex						
boys	14	76	24	11	70	30
girls	16	77	23	14	70	30
Age						
younger	14	75	25	12	72	28
older	15	77	23	11	74	26
Verbal IQ						
brighter	14	90	10	12	80	20
duller	15	65	35	10	68	32

determining the cut-off score for best diagnostic confidence. The corresponding point on the flatter graph below determines the misclassification rate at this cut-off point.

On the CDI the threshold or optimum cut-off score for identifying depression – as determined by the point of intersection for sensitivity and specificity – is 15 (rounded up from 14.5). Using this cut-off, the sensitivity and specificity rates are both 77% and the misclassification rate is 23%. On the DSRS, the optimum cut-off score is 11, which gives a specificity of 73%, sensitivity 75% and misclassification 25%. Other studies, using the DSRS, have reported broadly similar rates based on a cut-off score of 15 (Birleson *et al*, 1987) and 13 (Asarnow & Carlson, 1986).

To examine whether the efficiency of the CDI varied within the sample, the children were divided into the following subpopulations:

(a) male and female
(b) younger (12.7 years of age and below) and older (12.8 years of age and above)
(c) duller (verbal IQ of 90 and below) and brighter (verbal IQ of 91 and above).

Table 1 shows two clear findings. First, the misclassification rates on the CDI are similar for most of the subgroups. The exception is when the subgroup is defined by Verbal IQ; in this case the misclassification is significantly higher ($P<0.001$) for the duller group. A similar picture is found for the DSRS ($P<0.005$). Secondly, the cut-off scores for optimum discrimination, subgroup by subgroup, proved broadly similar, with the exception of separation according to sex. On the CDI, for girls the cut-off is 16 whereas for boys it is 14; on the DSRS, the scores are 14 for girls and 11 for boys.

The clinical interview included an assessment of anxiety as well as depressive symptoms. Clinical ratings of general anxiety were found to correlate significantly with the CDI ($r=0.49$) and DSRS ($r=0.57$) but there were no significant correlations with any specific types of anxiety.

The total scores of the two self-rating measures (CDI and DSRS) did not correlate significantly with age, sex and intelligence (vocabularly quotient) although this differed in relation to the optimum scores. Item analysis was also carried out; for brevity only significant levels are provided. Self-ratings of crying and tiredness proved more common among the girls ($P<0.01$), whereas poor peer relationships were more common among the boys ($P<0.01$). Sadness, lack of enjoyment, and guilt were more common among the older children ($P<0.02$) and somatic complaints and loneliness were more frequent among brighter children ($P<0.05$). One would expect to find some chance associations, but it is interesting to note that the findings are consistent with the notion that certain symptoms vary according to sex and stage of development (Rutter, 1986).

Discussion

Our results indicate a moderate but significant level of agreement between the CDI and the DSRS, and independent diagnosis, based on a Standardised Psychiatric Interview. However, although these two self-report measures are useful in screening children with affective disturbance, they are not precise diagnostic instruments. In brief, each of the screening measures was found to have a misclassification rate of about 25% in a referred clinic population. The optimum cut-off scores were 15 on the CDI and 11 on the DSRS. It is notable that both these cut-off scores are lower than those described by the authors of the CDI (Kovacs, 1981) and DSRS (Birleson, 1981). Such variations are probably due to differences in age range and population studied.

Girls tend to score higher than boys on depression scales (Tisher & Lang, 1983; Asarnow & Carlson, 1986) which is consistent with our findings. This suggests that for girls to be identified as probably suffering from depression on self-rating screening instruments, they require more symptoms or greater

severity of these than boys. It is possible that such differences may be a result of the manner in which girls and boys are socialised and in the way they express affect and distress (Izard & Schwartz, 1986; Gerde *et al*, 1988). Nevertheless, to date the designers of the self-rating depression instruments have not suggested different cut-off scores for boys and girls.

Our finding that optimum discrimination on both instruments was obtained on the brighter children is of interest. It has been suggested that within the prescribed age range for specific self-rating questionnaires, younger children are likely to interpret the statements in a more concrete manner leading to an excess of random responses and spurious results (Eysenck & Eysenck, 1975; Firth & Chaplin, 1987). Optimum discrimination on both instruments was obtained with brighter children, which suggests that the efficiency of the two instruments was more influenced by verbal intelligence than chronological age. In other words, within the age range of this sample, it was the duller child rather than the younger child who gave responses leading to higher rates of misclassification. Is this because duller, less verbally able children have a greater difficulty in differentiating between the meanings of the different statements on the self-rating measures or is it because they find it more difficult to introspect and, therefore, to describe their true thoughts and feelings in terms of the self-rating statements of the questionnaire?

Finally, the sex differences in relation to the optimum cut-off score suggest that the use of a single threshold or cut-off score for boys and girls may be unwise. Indeed the point was made some years ago (Saylor *et al*, 1984) that sex and age differences in relation to certain of the self-report measures on depression needed to be systematically addressed. A survey of the literature suggests that, by and large, that same argument still applies.

The self-report measures of depression were found to have disappointing sensitivity and specificity and this limits their clinical utility in a clinic population. However, this is not necessarily what might be found in the general population.

Previous studies have shown a relationship between childhood depression, conduct problems (Puig-Antich, 1982), and poor academic performance (Leon *et al*, 1980). However, in our study, self-ratings by children of depression (on the CDI and DSRS) were not associated with adverse ratings by teachers on behaviour or scholastic performance. Thus, either the children with depressive symptoms were functioning adequately, or any recent dysfunction on their behalf remained undetected by teachers. Unfortunately, the symptom of dysphoric mood was not included in the teacher scales and, therefore, we do not know if the teachers were aware of their pupils' distress. Such awareness by teachers as independent observers is only likely with availability of comprehensive teacher questionnaires geared more specifically to child depression.

The Conners Teacher Questionnaire simply measures specific aspects of observable behaviour that have to be rated as to presence and severity, and does not allow for exploration of the child's feelings and thoughts. Hence, sensitive discrimination between depressed and non-depressed children and identification of concurrent illness may be beyond the scope of a questionnaire administered by teachers.

*T. Fundudis, PhD MA, DipPsychother, CPsychol, FBPsS, *Top Grade Psychologist, Fleming Nuffield Unit, Newcastle upon Tyne*; T. P. Berney, MB, ChB, FRCPsych, *Consultant Psychiatrist, Fleming Nuffield Unit, Newcastle upon Tyne*; I. Kolvin, BA, MD, FRCPsych, DipPsych, *Professor of Child and Family Mental Health, Royal Free Hospital School of Medicine and Tavistock Clinic, London*; O. O. Famuyiwa, BM, MRCPsych, *formerly Research Fellow, Fleming Nuffield Unit, Newcastle upon Tyne*; L. Barrett, MB, BS, MRCPsych, *Consultant Child Psychiatrist, Queen Elizabeth Hospital, Gateshead*; S. Bhate, MB, BS, FRCPsych, *Consultant Child and Adolescent Psychiatrist, Newcastle General Hospital*; S. Tyrer, MB, BChir, FRCPsych, *Consultant Psychiatrist, Royal Victoria Infirmary, Newcastle upon Tyne*

*Correspondence