

## The assessment of lifetime psychopathology: a comparison of two interviewing styles

RICHARD HARRINGTON,<sup>1</sup> JONATHAN HILL, MICHAEL RUTTER, KAREN JOHN,  
HAZEL FUDGE, MARK ZOCCOLILLO AND MYRNA WEISSMAN

*From the Institute of Psychiatry, London*

**SYNOPSIS** A structured respondent-based version of the SADS-L interview was compared with a semi-structured investigator-based version of the same interview in terms of efficiency in assessing lifetime psychopathology. There was substantial agreement across most Research Diagnostic Criteria (RDC) categories, but there was less agreement on the rating of symptoms. The implications of these similarities and differences between the techniques are discussed.

### INTRODUCTION

It is now widely accepted that to achieve reliable diagnoses in psychiatric research it is necessary to use standardized interview techniques. Nonetheless, different views have been put forward about the degree of structure required in these interviews and about the style of questioning needed to obtain accurate and meaningful information about psychopathology. Thus, some authors have advocated semi-structured approaches in which there is both systematic coverage of all relevant areas and flexible probing of the subject's responses to determine whether the feature in question meets the *investigator's* criteria (Wing *et al.* 1967). In contrast, others have advocated more structured techniques in which the rating is determined by the *respondent's* replies to set questions (Robins *et al.* 1981).

There have been several recent comparisons between semi-structured and structured interviews (Farmer *et al.* 1987; Hesselbrock *et al.* 1982), and between structured interviews conducted by lay interviewers and standardized interviews conducted by clinicians (Anthony *et al.* 1985; Helzer *et al.* 1985). The focus of these studies, however, was the comparison between different instruments that were administered by interviewers with different clinical backgrounds,

rather than the examination of differences of interview technique.

The purpose of the present study is to compare two different interview techniques that are both based on the same instrument, namely, the Schedule for Affective Disorders and Schizophrenia Lifetime Version: SADS-L (Spitzer & Endicott, 1975). The study is part of a pilot study for a collaborative investigation of depressive disorders which is currently being carried out by a research group in the United Kingdom (UK) and a research group in the United States (US). The present study compares the flexible, investigator-based style of interviewing used by the UK group with the more structured, respondent-based style of interviewing used by the US investigators. One goal is to assess the amount of agreement between the two styles on items such as lifetime diagnosis, age at onset, and symptoms. Another goal is to investigate the reasons for disagreement, and to examine the strengths and weaknesses of the different styles.

### METHOD

#### Subjects

To compare the interviewing styles most effectively it was necessary to use a sample with a relatively high rate of psychiatric disorder, since assessment of agreement may be confounded by low base rates of disorder (Spitznagel & Helzer, 1985). In addition, since this study was a pilot for a family-genetic investigation of

<sup>1</sup> Address for correspondence: Dr R. C. Harrington, Department of Child and Adolescent Psychiatry, Institute of Psychiatry, De Crespigny Park, London SE5 8AF.

the adult relatives of child psychiatric patients, we wished to include subjects with the kinds of disorders that were expected in the main study. Accordingly, the 40 subjects were recruited from two sources: adult psychiatric out-patients ( $N = 18$ ) and parents of children attending the children's department ( $N = 22$ ). In both groups the subjects were as near as possible a consecutive series of clinic attenders. There were equal numbers of men and women, and their ages ranged from 24 to 60 years (mean age 42). Fourteen subjects had been psychiatric in-patients in the past, and a further 14 had had some form of out-patient treatment for psychiatric disorder.

#### Procedures

Each subject was interviewed twice (mean time between interviews 30 days, range 16–56 days), a different style being employed on each occasion by a different interviewer. Twenty subjects had the flexible style interview first and the structured style interview second; 20 had the interviews in the reverse order. Each of the two interviewers carried out an equal number of interviews with each style, and an equal number of first and second interviews. Subjects were randomly allocated to these conditions. To minimize loss of information between interviews that might occur because subjects were not aware of the purpose of the second interview, subjects were reminded again of the purpose of the study on the second occasion. All the interviews were tape-recorded. Ratings were made independently.

Because the 'investigator-based' technique required an interviewer with some knowledge of psychiatric concepts and because we wished to avoid the possibility, raised in previous research (Robins *et al.* 1982), that informants say more or less to an interviewer simply because of his professional status, both interviewers were clinicians.

Training for the structured 'respondent-based' interview was carried out in several stages and involved completion of a training package (containing case vignettes and videotaped interviews) and visits by a US interviewer (K.J.) to the UK group. Throughout the study the structured SADS-I was completed in the order set out on the form, and no attempt was made to vary the order of the interview according to the subject's responses. The overview was

restricted to establishing a few details about treatment and hospitalizations and there were few open questions. Subjects were asked directly about symptoms but not encouraged to expand or to give examples. Their replies to direct questions were accepted without cross-questioning but, when necessary, questions were repeated to ensure that the subject had understood them. Symptoms were rated simply as present or absent.

In contrast, the flexible 'investigator-based' style began with an open overview, intended to give the floor to the subject at an early point and to identify those disorders that particularly troubled him. The interview was flexible to the extent that the subject's account in the overview determined which areas were then examined first, and his verbal cues were followed up. Questioning on symptoms started with open questions and then proceeded to direct questions. If the subject gave a positive reply to the initial question he was then asked to give a description of the symptom. Severity was rated on a 0–3 point scale after cross-questioning (0 = absent; 1 = not significant; 2 = definitely present, moderate severity; 3 = definitely present, severe). For this purpose, supplementary probes were added to the schedule to guide cross-questioning, and a rating manual was devised which contained a detailed definition of each symptom and guidelines on the rating of severity.

To see if the interviewers remained consistent in their use of the two styles during the study, another rater listened to twelve randomly selected tape-recorded interviews (six in each style, six by each interviewer). The rater, who did not know beforehand which style was being used or that the sample contained six interviews in each style, was asked to guess the style of interviewing. All twelve guesses were correct.

Both interviews were divided into sections that were preceded by screening questions. These sections were entered if the subject made a positive response to one of the screening questions and they covered the information necessary to make diagnoses based on the Research Diagnostic Criteria (RDC: Spitzer *et al.* 1975), with the modifications described by Mazure & Gershon (1979). A non-hierarchical method of diagnosis was employed, so that it was possible for a subject to have more than one diagnosis. Twenty-three RDC diagnoses were

assessed by both interviews, of which 20 were diagnosed at least once in this study. Seven diagnoses were so rare as to render questionable the reporting of statistics for them individually, and they are discussed only briefly in the text.

### Analyses

Since many subjects had multiple diagnoses, agreement between the interviewing styles was assessed by a proportional overlap procedure (Mezzich *et al.* 1981) which defines proportion of agreement between two diagnostic formulations as the ratio of number of agreements between specific categories over the number of possible agreements. Kappa was computed from this by the method described by Mezzich *et al.* To identify disorders that were particularly problematical, Kappa coefficients were then calculated for each diagnosis separately. Kappa coefficients were also used to assess agreement on other dichotomous items such as symptoms (to compare agreement about symptoms, symptom ratings in the flexible interview were dichotomized: '0' and '1' ratings into 'absent'; '2' and '3' ratings into 'present'). Agreement on quantitative items such as age at onset was examined by computing intraclass correlation coefficients (Bartko, 1966).

Paired *t* tests were used to examine whether there were significant differences between the two interview techniques on quantitative variables such as the number of diagnoses per subject, or the subjects' preference for one or other interview. To take into account the simultaneous effects of interviewing style, interviewer, and occasion on these variables, these effects were examined within a regression model. The results were not materially altered by this analysis.

## RESULTS

### Inter-rater reliability of the investigator-based style

To assess the inter-rater reliability of the investigator-based style, the audiotapes of 20 interviews from the study were randomly selected, and rated independently by a psychiatrist who had not been involved in the interviewing (M.Z.). The Kappa for multiple diagnoses was 0.8. Kappa coefficients for separate diagnoses ranged from 0.63 for other psychiatric disorder

Table 1. Agreement between structured and flexible interviews for specific lifetime diagnoses

Diagnosis	Base rate (%)	Kappa
Major depression	35	0.84
Alcohol abuse	25	0.94
Never mentally ill	17.5	0.71
Minor depression	17.5	0.55
Generalized anxiety disorder	17.5	0.60
Anti-social personality	17.5	1
Hypomania	12.5	0.59
Drug abuse	12.5	0.89
Phobia	12.5	0.80
Bereavement	12.5	0.72
Briquet's disorder	10	0.87
Panic disorder	5	0.64
Other psychiatric disorder	5	0.22

to 0.89 for major depression, with a mean value for the eight diagnoses assessed of 0.81. These results are similar to those reported in previous studies of the inter-rater reliability of the SAIDS-I, in samples of psychiatric patients (Spitzer *et al.* 1978) and in samples consisting of patients, their relatives, and medical controls (Mazure & Gershon, 1979).

### Agreement on lifetime diagnosis

To compute the overall Kappa for multiple diagnoses, all diagnoses made on the subject were included (in both interviews the number of diagnoses per subject ranged from 0 to 8; mode = 2). The overall Kappa was 0.62 ( $t$  6.32,  $P < 0.001$ ).

Table 1 presents results concerning agreement on the most common RDC diagnoses. Base rates were determined by using the consensus ratings and they add up to more than 100% because many subjects had multiple diagnoses. There was good agreement on many diagnostic categories but agreement was lower about the diagnoses of minor depression, hypomania, and other psychiatric disorder. In the modified Research Diagnostic Criteria (Mazure & Gershon, 1979) both minor depression and hypomania can be diagnosed on the basis of symptoms alone and do not require another criterion such as social impairment. Disagreements about 'other psychiatric disorders' (psychiatric disorders in which there was social impairment or help seeking but which did not meet the diagnostic criteria for the other categories) are discussed further below.

Agreement on rare diagnoses was more

variable. There was perfect agreement about the two diagnoses of mania. On the other hand, there was no agreement about the two diagnoses of schizophrenia (one diagnosis was made by each interview), about the two diagnoses of post-partum psychiatric disturbance, or about the one diagnosis of organic brain syndrome.

To examine whether one style of interviewing produced more diagnoses than the other, the number of diagnoses per subject was compared by a paired *t* test. The difference was not significant. We then examined each diagnosis separately, but there was no significant bias between the interviews on any single disorder except that 'other psychiatric disorders' were more frequently diagnosed by the investigator-based style ( $P < 0.05$ ). This category included disorders like repeated deliberate self-harm and multiple personality. There were no significant interviewer or order effects.

#### Agreement on symptoms, age at onset, and other items

Agreement on symptoms was generally lower than that found for diagnoses. For example, there were twelve questions on specific symptoms that preceded the sections of the interviews (and were therefore examined in every subject) and which were examined by both interviews. Table 2 presents the base rates and Kappa coefficients for these 'core' symptoms. Kappa ranged from 0.37 for obsessions to 0.81 for depressed mood. In only four instances did Kappa exceed 0.7.

The findings were similar for symptoms that

Table 2. Agreement between structured and flexible interviews for core symptoms (ever) examined in both interviews

Symptom	Base rate (%)	Kappa
Depressed mood	82	0.81
Irritable mood	25	0.69
Crying	18	0.51
High mood	18	0.48
Excitable mood	15	0.72
Panic attack	20	0.65
Anxiety	38	0.53
Obsessions	10	0.37
Phobic avoidance	38	0.50
Suicidal attempt	28	0.80
Dramatic medical history	18	0.80
Others objected to drinking	18	0.58

were contained within the sections of the interviews. For instance, in the 30 cases in which the 'depression' section was completed in both interviews, Kappa coefficients on separate symptoms ( $N = 20$ ) ranged from 0.12 for the symptom of pessimism to 0.93 for weight loss, with an average Kappa value of 0.53. Only two of these symptoms had Kappa values greater than 0.7.

In contrast, there was higher agreement on items indicating the severity of disorder (items that are essential for many RDC categories). Thus, for social impairment, the Kappa value was 0.84 and for the ever taking psychotropic drugs variable, Kappa was 0.91.

In general, the structured style scored a greater number of positive symptoms than the flexible style. Thus, nine out of the 12 core symptoms were rated more frequently by the structured style and 16 out of the 20 symptoms in the depression section were rated more frequently by the structured style ( $\chi^2 7.2$ ,  $P < 0.01$ ).

There was good agreement on the age at onset of the earliest diagnosis ( $r = 0.88$ ), and on the age at onset of specific diagnoses such as major depression ( $r = 0.98$ ), minor depression ( $r = 0.76$ ), and anxiety disorder ( $r = 0.93$ ). Dating of other characteristics, such as age at first outpatient care ( $r = 0.81$ ) and age at first hospital care ( $r = 0.95$ ), was also reliable.

#### Reasons for disagreement and the effects of probing

At the end of the study the interviewers reviewed each instance of diagnostic disagreement to determine a consensus rating on the reason for that disagreement.

There was a total of 41 disagreements about RDC diagnoses. Sixty-three per cent were attributed to the informants' inconsistencies in that they had simply said 'no' to a specific question on one occasion and 'yes' on another. For example, in one instance the subject had told one interviewer that schizophreniform symptoms had overlapped with the symptoms of major depression, and a diagnosis of schizoaffective disorder was made. He told the other interviewer that his schizophreniform symptoms did not overlap with depression and a diagnosis of schizophrenia was made. In many of these instances it seemed that disagreement had resulted from difficulties the subject had had in

accurate recall. However, our present sample was too small for us to consider covariates such as the time interval between the interview and the episode in question.

In 15% of disagreements the same information was obtained but it was interpreted differently because of inconsistencies on the part of the interviewers or ambiguities in the diagnostic criteria. In one instance disagreement was attributed to a change of the subject's mental state between interviews.

Twenty per cent of diagnostic disagreements were attributed to differences in interview technique. Of these, there were four instances in which the following up of the informant's verbal cues during the flexible interview led to information that changed the diagnosis. In one case, for example, the interviewer was able to obtain enough information to make a diagnosis of schizophrenia by encouraging the subject to expand on her bizarre comments about the radio.

Another important difference between the interview techniques was the extent to which the informants' replies to direct questions were probed to determine whether or not the feature met the definition set out in the manual. This supplementary probing accounted for only four disagreements about diagnosis. However, as we noted above, many symptoms were rated more frequently by the structured style and consequently a high proportion of disagreements about symptoms were due to symptoms being rated as 'present' by the structured style and 'absent' ('not present' or 'not significant') by the flexible and probing style. Furthermore, for some symptoms more than half of this type of disagreement was due to the symptom being rated 'not significant' by the probing style of interviewing (indicating that the subject had replied positively to the direct question in both interviews but that further probing during the flexible interview had shown that the symptom was not significant), rather than 'absent' by the flexible style (indicating that disagreement was due to the subject replying negatively to the direct question in the flexible interview). This occurred in five core symptoms (irritability, crying, high mood, anxiety, obsessions) and in five symptoms that were contained within the depression section (brooding, pessimism, feeling inadequate or resentful, physical complaints).

#### Informants' perceptions of the interviews

Immediately after each interview the informants were given a questionnaire (Rutter & Cox, 1981) regarding their feelings about the interview. After the second interview they were asked by another rater (H.F.), who did not know which interview had been administered, which interview they preferred in relation to seven aspects (see Cox *et al.* 1981), such as whether they felt relaxed or tense, and whether the interviewer had given them a chance to speak. There were no significant differences between the two interviewing styles on any of the questionnaire scales, or on any of the items assessed by the rater, except that the flexible style was perceived as taking longer than the structured style ( $t = 4.55$ ,  $P < 0.001$ ). This was consistent with the interview times recorded by the interviewers (the mean times were 93 and 60 minutes respectively). The interviewers were perceived as being less interested on the second occasion than on the first ( $t = 2.12$ ,  $P < 0.05$ ), but there was an overall preference for the second occasion ( $t = 2.1$ ,  $P < 0.05$ ).

#### DISCUSSION

This study compared a flexible investigator-based interview technique with a more structured respondent-based interview technique in terms of their efficiency in assessing 'lifetime' psychopathology. It should be borne in mind that although the structured style had many of the features of a structured scheduled interview, in practice it was not as highly structured as some schedules that assess lifetime psychopathology, such as the Diagnostic Interview Schedule (Robins *et al.* 1981). Furthermore, in this study both interviews were administered by clinicians. Thus, the findings may not be directly applicable to the issue of the reliability between interviews administered by clinicians and highly structured interviews administered by interviewers with no clinical experience (Spitzer, 1984; Klerman, 1985).

Nonetheless, the findings were striking in showing not only substantial agreement between the structured and semi-structured styles but also important areas of disagreement. Thus, diagnostic agreement across most RDC categories was as high as that found in test-retest

studies of the SADS-L (see e.g. Andreasen *et al.* 1981; Mazure & Gershon 1979) and our findings of lower agreement on minor depression and hypomania were also found in this work. On the other hand, agreement on individual items was generally lower than agreement on diagnosis. This suggests that this high diagnostic agreement was partly due to the robustness of the diagnostic categories. Indeed, supplementary probing during the flexible interview accounted for a high proportion of disagreement on some items. This was particularly the case for symptoms that are part of normal emotional functioning, such as crying and obsessional phenomena, which are morbid only in their intensity and persistence. In these instances supplementary probing was often required to clarify the meaning of initial answers and to establish whether a symptom had persisted for the length of time required by the definition.

The finding that the two approaches differed in the extent to which they elicited this type of symptom suggests that there are uncertainties involved in diagnoses based on the number of such symptoms only. There is likely to be better diagnostic agreement if greater attention is paid to more serious psychopathological phenomena, and/or if there is reliance on substantial social impairment, which proved to be a reliable item in this study. This is likely to be especially important for studies of relatives of index patients, in which a large percentage of subjects have symptomatology near the threshold for diagnosis (Coryell *et al.* 1981). Our present sample, however, was too small for us to consider covariates such as severity although it is interesting that overall diagnostic agreement was higher in the sub-sample of psychiatric outpatients (Kappa = 0.69) than in the sub-sample of parents (Kappa = 0.57).

The features discussed so far refer to items that were routinely assessed by both interviews. The structured style tended to score more of these items as positive but there was some evidence that the flexible style identified more items that were not directly examined by both interviews. Non-RDC diagnoses were more frequently identified by the flexible style of interviewing. This finding suggests that an open style of questioning is better able to identify psychiatric disorders outside the range of standard enquiry. Conversely, this style of question-

ing added to the length of time that the interview took and some subjects said that they preferred the more rapid, more structured style. In this regard, it should be noted that other preferred to be given the chance to tell the story in their own words and that the majority of subjects found both styles of interviewing acceptable.

Another difference between the interview techniques concerned the way information was recorded. The organization of the SADS-L was designed to elicit information necessary for making diagnoses using the RDC and this has the advantage that a standardized diagnosis can be made immediately after the evaluation (Endicott & Spitzer, 1978). There is the disadvantage, however, that the schedule is tied to these diagnostic conventions and it is therefore difficult to use the information for other purposes. An advantage of an approach that emphasizes the recording of events and behaviours, rather than the recording of translations of specified criteria, is that it provides information that can be used more flexibly. For example, in this study the information on work history which was collected by the interviewer using the flexible style was used not only for the RDC diagnosis of anti-social personality but also to provide data for an assessment of the subject's level of adaptive functioning throughout adult life (Quinton *et al.* 1984). This procedure did not adversely affect agreement on the RDC diagnosis of antisocial personality.

To summarize, there was adequate agreement between the two interview techniques across many diagnoses but there was less agreement about the rating of symptoms. Differences of interview technique accounted for a substantial minority of diagnostic disagreements, and for a higher proportion of disagreement about some symptoms. It is notable, however, that many instances of diagnostic disagreement were due to the informant giving different accounts on the two occasions. This finding underlines the difficulty that many subjects have when they are asked to remember symptoms that may have occurred at any time during their lives but may no longer be present. In this regard, it should be emphasized that raters in actual practice have more data available than was provided in the design of this study. Thus, raters usually obtain

multiple informant accounts and patient records if these are available, and make best estimate lifetime diagnoses making use of all available information (Leckman *et al.* 1982). Nonetheless, direct information from the subject is an important source of information and the quality of this information is highly dependent on the interviewer's skills in helping the subject with the challenging task of recalling past events.

The research was supported by a grant from the MacArthur Foundation Network on Risk and Protective Factors in Mental Disorders. Dr Hill was supported by the Medical Research Council and Dr Zoecolillo by a Fellowship from the National Institute of Mental Health.

Thanks are due to the clinicians of the Maudsley Hospital for their active collaboration and to Graham Dunn and Brigitte Prusoff for their suggestions on the statistical methods.

## REFERENCES

- Andresen, N. C., Grove, W. M., Shapiro, R. W., Keller, M. B., Hirschfeld, R. M. A. & McDonald-Scott, P. (1981). Reliability of lifetime psychiatric diagnosis: a multi-center collaborative perspective. *Archives of General Psychiatry* 38, 400-405.
- Anthony, J. C., Folstein, M., Romanoski, A. J., Von Koff, M. R., Nestadt, G. R., Chahal, R., Merchant, A., Brown, C. H., Shapiro, S., Kramer, M. & Guenberg, E. M. (1985). Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis: experience in Eastern Baltimore. *Archives of General Psychiatry* 42, 667-675.
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19, 3-11.
- Coryell, W., Winokur, G. & Andreasen, N. C. (1981). Effect of case definition on affective disorder rates. *American Journal of Psychiatry* 138, 1106-1109.
- Cox, A., Rutter, M. & Holbrook, D. (1981). Psychiatric interviewing techniques. V. Experimental study: eliciting factual information. *British Journal of Psychiatry* 139, 29-37.
- Endicott, J. & Spitzer, R. L. (1978). A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry* 35, 837-844.
- Farmer, A., Katz, R., McGuffin, P. & Bebbington, P. (1987). A comparison between the Composite International Diagnostic Interview and the Present State Examination. *Archives of General Psychiatry*. (In the press.)
- Helzer, J. E., Robins, E. N., McEvoy, L. T., Spitznagel, E. L., Stoltzman, R. K., Farmer, A. & Brockington, J. F. (1985). A comparison of clinical and Diagnostic Interview Schedule diagnoses: physician reexamination of lay-interviewed cases in the general population. *Archives of General Psychiatry* 42, 657-666.
- Hesselbrock, V., Stabenau, J., Hesselbrock, M., Mirkin, P. & Meyer, R. (1982). A comparison of two interview schedules: the Schedule for Affective Disorders and Schizophrenia Lifetime and the National Institute for Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry* 39, 674-677.
- Klerman, G. L. (1985). Diagnosis of psychiatric disorders in epidemiologic field studies. *Archives of General Psychiatry* 42, 723-724.
- Leckman, J. F., Sholomskas, D., Thompson, W. D., Belanger, A. & Weissman, M. M. (1982). Best estimate of lifetime psychiatric diagnosis: a methodological study. *Archives of General Psychiatry* 39, 879-883.
- Mazure, C. & Gershon, E. S. (1979). Blindness and reliability in lifetime psychiatric diagnosis. *Archives of General Psychiatry* 36, 521-525.
- Mezzich, J. E., Kraemer, H. C., Worthington, D. R. L. & Collman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of Psychiatric Research* 16, 29-39.
- Quinton, D., Rutter, M. & Ladle, C. (1984). Institutional rearing, parenting difficulties and marital support. *Psychological Medicine* 14, 107-124.
- Robins, E. N., Helzer, J. E., Croughan, J. & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Archives of General Psychiatry* 38, 381-389.
- Robins, E. N., Helzer, J. E., Ratcliff, K. S. & Seyfried, W. (1982). Validity of the Diagnostic Interview Schedule, Version II (DISM-III) diagnoses. *Psychological Medicine* 12, 855-870.
- Rutter, M. & Cox, A. (1981). Psychiatric interviewing techniques: I. Methods and measures. *British Journal of Psychiatry* 138, 273-282.
- Spitzer, R. L. (1984). Psychiatric Diagnosis: are clinicians still necessary? In *Psychotherapy Research: Where Are We And Where Should We Go?* (ed. J. R. W. Williams & R. L. Spitzer), pp. 273-288. Guilford Press: New York.
- Spitzer, R. L. & Endicott, J. (1975). *Schedule for Affective Disorders and Schizophrenia - Lifetime Version*. Biometrics Research: New York.
- Spitzer, R. L., Endicott, J. & Robins, E. (1975). *Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders*. Biometrics Research: New York.
- Spitzer, R. L., Endicott, J. & Robins, E. (1978). Research Diagnostic Criteria: rationale and reliability. *Archives of General Psychiatry* 35, 773-782.
- Spitznagel, E. L. & Helzer, J. E. (1985). A proposed solution to the base rate problem in the Kappa statistic. *Archives of General Psychiatry* 42, 725-728.
- Wing, J. K., Birley, J. L. T., Cooper, J. E., Graham, P. & Isaacs, A. D. (1967). Reliability of a procedure for measuring and classifying 'present psychiatric state'. *British Journal of Psychiatry* 113, 499-515.