Psychological Medicine, 2001, **31**, 1001–1013. © 2001 Cambridge University Press
DOI: 10.1017/S003329170105418 Printed in the United Kingdom

A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN)[1]

T. S. BRUGHA,[2] R. JENKINS, N. TAUB, H. MELTZER AND P. E. BEBBINGTON

From the Department of Psychiatry and Department of Epidemiology and Public Health, University of Leicester; and WHO Collaborating Centre, Institute of Psychiatry, Social Survey Division, Office for National Statistics and Department of Psychiatry and Behavioural Sciences, University College, London

ABSTRACT

Background. In psychiatric surveys of the general population, there has been considerable discrepancy between diagnoses obtained by fully structured interviews and those established by systematic semi-structured clinical evaluation. The Composite International Diagnostic Interview (CIDI) is an example of the first type of interview widely used in general population surveys. We compared its performance in diagnosing current depressive and anxiety disorders with the Schedules for Clinical Assessment in Neuropsychiatry (SCAN), a semi-structured diagnostic interview administered by clinically trained interviewers.

Methods. Household addresses in Leicestershire, UK, were randomly sampled and 860 adults were screened with the Revised Clinical Interview Schedule. Adults with too few symptoms to fulfil diagnostic criteria for study disorders were excluded to increase the proportion re-interviewed who met such criteria. Repeat diagnostic interviews with the CIDI and SCAN, ordered randomly, were sought from eligible screen positive respondents. Recalibrated CIDI prevalence estimates were derived from the SCAN classification using Bayesian statistics.

Results. Concordance ranged between 'poor' and 'fair' across almost all types of study disorders, and for co-morbidity. Concordance was somewhat better for severity of depression and when lower diagnostic thresholds were used for depression. Interview order effects were suggested with lower concordance when CIDI followed SCAN. Recalibration reduced the prevalence of depressive or anxiety disorder from 9·0 to 6·2%.

Conclusions. Community psychiatric surveys using structured diagnostic interview data must be interpreted cautiously. They should include an element of clinical re-appraisal so findings can be adjusted for estimation differences between fully structured and clinical assessments.

INTRODUCTION

The rational allocation of mental health care resources (Brugha *et al.* 1997; Murray & Lopez, 1997) is assisted by valid information on the prevalence of disorder. However, there has been scepticism about the validity of fully structured diagnostic psychiatric interviews, notably among pubic health physicians and health policy makers (Bartlett & Coles, 1998; Leeman, 1998). This has been increased by findings of substantial differences in the prevalence of depression in the USA in two large-scale surveys of psychiatric disorder employing such methods (Regier *et al.* 1998).

[1] Preliminary findings from this study were communicated at: the Association of European Psychiatrists, Cambridge, April 1996; the WPA Section of Epidemiology and Community Psychiatry, Sydney, Australia, 19 October 1997; and the International Association of Epidemiologists Congress (IEA99), Florence, Italy, September 1999.

[2] Address for correspondence: Professor T. S. Brugha, University of Leicester, Department of Psychiatry, Section of Social and Epidemiological Psychiatry, Brandon Mental Health Unit, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW.

We have argued at length that as most psychiatric diagnoses have no satisfactory external validation, for instance through detectable pathophysiology, diagnostic instruments should be validated against the systematic observation of the syndrome in a way closest to the conceptualization underlying it (Brugha *et al.* 1999*a*). This evaluation should be done by clinically trained interviewers.

One example of a fully structured interview is the Diagnostic Interview Schedule (DIS) (Robins *et al.* 1981). In general population surveys, the DIS, used by trained interviewers, yielded poor concordance, both with clinician administered DSM-III checklists (Helzer *et al.* 1985) and with a present state Standardized Psychiatric Examination (Anthony *et al.* 1985; Romanoski *et al.* 1988). The results from such validation exercises (Anthony *et al.* 1985; Helzer *et al.* 1985; Kessler *et al.* 1998) may have been compromised by their incorporation in large-scale surveys with different aims (Brugha *et al.* 1999*a*). Moreover, the clinician assessments were always conducted after the fully structured interview assessments, possibly introducing bias due to order effects. Delays between the main survey interview and clinician re-appraisal may also have reduced concordance (McLeod *et al.* 1990).

The Composite International Diagnostic Interview (CIDI) (World Health Organization 1993*a*) was developed as an improvement on the DIS, whose features it largely shares (Robins *et al.* 1988). Subsequently, extensive efforts were made to improve the comprehension, motivation and recall of respondents, and concordance with clinical evaluations was better with this instrument (Kessler *et al.* 1998). However, these evaluations employed non-blinded clinically trained interviewers, completing single diagnostic modules of the Structured Clinical Interview Schedule for DSM-IV (Spitzer *et al.* 1992). The validity of the CIDI as an actual case finding instrument in community surveys therefore remains to be evaluated. This requirement is more easily stated than achieved.

In the general population, psychiatric symptoms may be transient and relatively mild. Sources of diagnostic invalidity in community surveys may include respondents' lack of understanding of their task, or of the question asked, and a lack of ability and willingness to carry out the task (Biemer *et al.* 1991; Kessler *et al.* 1998; Turner *et al.* 1998). Diagnostic concordance has sometimes been better in clinical settings (Wittchen, 1994), and this may arise because patients may learn from their physician or psychologist the correct meaning of terms such as anxiety, phobia, panic and obsession (Brugha *et al.* 1999*a*).

Assessing correctly the presence of psychopathological phenomena requires expert judgement (Brugha *et al.* 1999*a*). Trainees learn to elicit and assess respondents' individual descriptions of mental states, and to distinguish pathological from normal phenomena. These two skills must be clearly specified and taught (Brugha *et al.* 1999*a*). Standardized, semi-structured interviews systematize this clinical process (Wing *et al.* 1990; Spitzer *et al.* 1992) and we would argue in theory, therefore, that they provide more valid assessments than wholly structured interviews (Brugha *et al.* 1999*a*). Nevertheless, measurement error, for example arising from inter-interviewer variability in rating thresholds, may undermine validity (Bromet *et al.* 1986; Wittchen *et al.* 1999).

The interpretation of data in mental health community surveys derived from fully structured interviews may be assisted by assessing overall survey error (Kruskal, 1991), for which direct clinical assessment has been advocated (Brugha *et al.* 1999*a*). According to Kruskal 'our motivations for attacking errors … of measurements are mainly to gain improved understanding of the process, thus making better decisions based on the measurements …' (Kruskal, 1991). After the completion of the British National Psychiatric Morbidity Survey (Jenkins *et al.* 1997*a*), a separate survey was conducted with this in mind (Brugha *et al.* 1999*b*). Ratings on SCAN, a semi-structured diagnostic interview administered by specially trained interviewers (Wing *et al.* 1990), were compared with data from a fully structured diagnostic interview developed in Great Britain (Lewis *et al.* 1992), and from the CIDI (Robins *et al.* 1988). The feasibility of prevalence estimates derived from the fully structured diagnostic interviews was examined by using SCAN as the reference measure.

METHOD

Sample

Two thousand five hundred Postcode Address File delivery points (Wilson & Elliot, 1987) were randomly sampled in urban, suburban and rural Leicestershire, England, an area chosen because it has socio-economic characteristics representative of Great Britain as a whole (Brugha *et al.* 1999*b*). Addresses were allocated randomly to interviewers, who then randomly sampled one eligible adult within each household (Kish, 1965). Respondents had to be aged 16 to 64 years and capable of an interview in English. Adults normally resident in institutions or elsewhere were excluded.

Design

A two-phase survey design was used (Pickles *et al.* 1995) (Fig. 1). In phase one, eligible adults were screened with the Revised Clinical Interview Schedule (CIS-R) (Lewis *et al.* 1992). The frequency of ICD-10 diagnoses at each CIS-R score level was determined using national data (Meltzer *et al.* 1995). Respondents with scores below 8 on the CIS-R were not considered further because they were very unlikely to meet diagnostic criteria on the follow-up instrument (Brugha *et al.* 1999*b*; Meltzer *et al.* 1995). Using random numbers, subjects who were screen positive were allocated either to have a SCAN phase two diagnostic interview or a CIDI diagnostic interview. The second phase two interview, using the remaining diagnostic measure, was then to be completed within 2 weeks (Fig. 1).

If the population value of the primary concordance estimator kappa (Cohen, 1968) is 0·4, between 150 and 200 successful pairs of interviews would be required to test the null hypothesis of a zero value for kappa at the 5% significance level with 80% power.

Measures

The CIDI and the SCAN were developed in parallel under WHO auspices (Robins *et al.* 1988; Wing *et al.* 1990), and are widely endorsed techniques for case identification. The CIDI is a fully structured diagnostic interview (World Health Organization, 1993). It employs precisely worded questions that cannot be rephrased but may be repeated.

The 10th version of the semi-structured, Present State Examination, is the central component of the SCAN Version 1 (Wing *et al.* 1998; World Health Organization Division of Mental Health, 1992). Every symptom in SCAN is defined in detail (Wing *et al.* 1990). Suggested wording is provided for eliciting each SCAN symptom. However, interviewers must probe further until satisfied with the information obtained, because it is they who must decide if symptom definitions are fulfilled.

Clinical knowledge is not required for the CIDI, but must be obtained in order to use SCAN, although lay survey interviewers have been trained to use it reliably in a clinical population (Brugha *et al.* 1999*a*).

Interviewers and procedures

Fourteen interviewers were recruited. Twelve were non-medical interviewers, including five university psychology graduates. All 14 interviewers were trained by the Office of Population Censuses and Surveys (OPCS) (National Statistics) in structured interviewing, including the CIDI, and in the survey sampling techniques used in the British National Surveys (Meltzer *et al.* 1995). They were assisted by an experienced CIDI trainer. Two interviewers were physicians, with at least 3 years postgraduate experience in clinical psychiatry, who completed a standard 5-day course in the use of the SCAN (Üstün *et al.* 1998). They undertook further training and pilot SCAN interviews until the reliability of their ratings was fully concordant with senior SCAN trainers rating interviews alongside them.

The study aim required that direct comparisons of diagnostic outputs should reveal differences in the method of interview assessment (Brugha *et al.* 1999*a*), not differences in diagnostic procedures or differences in the type of interviewer. Interviewers introduced themselves to respondents as University of Leicester survey interviewers (not as psychologists or physicians). Both CIDI and SCAN were computer assisted: automated cut-offs minimized unnecessary or inappropriate questions (World Health Organization, 1993*a*, *b*; Der *et al.* 1998). Both interviews were designed to collect data required by diagnostic criteria, and employ identical classification rules (World Health Organization, 1993*a*, *b*).

Interviewers visited respondents' homes, and sought consent for the CIDI and SCAN interviews from screen positive respondents (Brugha *et al.* 1999*b*). Only 14 days were permitted between administration of the CIDI and SCAN. The two SCAN interviewers were allocated cases according to their availability (non-randomly). All interviewers were blind to the results of other interviews on the same respondent.

Statistical analysis

Sociodemographic and clinical factors potentially associated with non-completion of follow-up interviews, were assessed using logistic regression models (Everitt & Der, 1996). Because we did not interview screen negatives, the findings have not been weighted for design or non-response.

ICD-10 algorithms developed by WHO for SCAN and CIDI (World Health Organization, 1993*a*, *b*; Der *et al.* 1998) were used to classify respondents into ICD-10 anxiety and depression categories (F32 to F42) within the past month (Tables 1 and 2). Diagnoses were generated non-hierarchically; thus subjects with more than one diagnosis are classified according to each disorder in the analysis. Different disorders were combined into broader categories, including a final 'catch all' category for respondents meeting criteria for any study disorder (Tables 1 and 2). Co-morbidity was defined as the co-occurrence of at least two separate study disorders.

Using SCAN as the reference instrument, we calculated kappa (Cohen, 1968; Landis & Koch, 1977), sensitivity, specificity and the proportion of subjects for whom a positive diagnosis was recorded by either observer in which both agreed (the Index of Agreement) (Cicchetti & Feinstein, 1990), with 95% confidence intervals.

Sensitivity analyses

Rigidly applied binary classifications (e.g. 'depressed or not') might conceal better agreement when different severity levels are considered. Kappa varies with the frequency of observations (Cicchetti & Feinstein, 1990). Raising or reducing threshold for depression (by requiring fewer criteria for a diagnosis) allows this to be evaluated. Although developed in parallel under WHO auspices, discrepancies may yet exist between published CIDI and SCAN algorithms for ICD-10, thus introducing classification error (Marcus & Robins, 1998). For these reasons a new depression algorithm was constructed common to both interviews, based on all the ICD-10 criteria for F32 to F33 (World Health Organization, 1993*a*, *b*). Thus, concordance was re-evaluated for: the diagnosis of depressive disorder; a total score for depression; and a range of thresholds representing varying numbers of depression criteria (Table 3).

Our assumption that the second stage screen misses very few cases is well supported by evidence (Brugha *et al.* 1999*b*). Nevertheless, concordance analyses were carried out in which both zero and complete concordance in such 'missed cases' was simulated. This analysis would tell us what concordance would have been if the unassessed cases had yielded either perfect agreement or no agreement.

As a check for variation between the two SCAN interviewers' data, key analyses were repeated after dividing the total sample into two subgroups A and B defined by SCAN interviewer.

Recalibrated prevalence estimation

To obtain recalibrated prevalence estimates, we simulated a two-stage survey of 10000 adults with CIDI used in the first stage and SCAN in the second. A simulated rate of 900 cases (any study depressive or anxiety disorder using CIDI) was used. This rate is close to that actually found in the present general population survey in Leicestershire (Brugha *et al.* 1999*c*). The CIDI prevalence rate was recalibrated by applying the estimated positive predictive value (proportion CIDI positive cases who are SCAN positive) and negative predictive value (proportion CIDI negative cases who are SCAN negative).

In order to estimate the precision of the recalibrated prevalence rate, confidence intervals were obtained from a Bayesian graphical model using BUGS computer software (Spiegelhalter *et al.* 1995). This enabled us to take account of assumptions made about the distribution of a variable before new empirical data are considered. Thus, it allowed for the different sampling procedures. For example, the

Table 1. *Concordance, true negatives, false positives, false negatives and true positives of SCAN and CIDI-Auto ICD-*10 *diagnoses present in the month before interview in second-stage screen positive respondents in the general population* (N = 172)

		SCAN/CIDI					
		−/−	−/+	+/−	+/+		
Diagnosis	ICD10 Code(s)	Both interviews negative	Only CIDI positive	Only SCAN positive	Both interviews positive	kappa*	95% CI
Depressive disorders							
Any depressive episode	F32.00, F32.01, F32.10, F32.11, F32.20	150	16	3	3	0·20	−0·02 to 0·42
Depressive episode or disorder	Any F32 or any F33	144	22	3	3	0·15	−0.04 to 0·33
Dysthymia†	F34.1	158	5	3	4	0·48	0·17 to 0·79
Any depression above (not remission)	Any F32 or any F33 or F34.1	141	19	3	9	0·39	0·19 to 0·59
Anxiety disorders							
Any agoraphobic anxiety disorder	F40.00, F40.01	156	12	1	3	0·29	0·02 to 0·56
Agoraphobia with panic disorder	F40.01	166	4	1	1	0·27	−0·17 to 0·71
Social phobias	F40.10	161	8	0	3	0·41	0·09 to 0·73
Specific/isolated phobia	F40.20	130	21	9	12	0·35	0·17 to 0·53
Any phobic anxiety above	F40.00, F40.01, F40.10, F40.20	120	27	8	17	0·38	0·22 to 0·54
Panic disorder	F41.00, F41.01	145	13	9	5	0·24	0·02 to 0·46
Generalized anxiety disorder (GAD)	F41.10, F41.11	152	17	3	0	−0·03	−0·06 to 0·00
Any non-phobic anxiety disorder	Any F41 (Panic or GAD) or F41.8	132	23	9	8	0·24	0·05 to 0·42
Any phobia, panic, or GAD	Any F40 or F41	103	33	12	24	0·35	0·20 to 0·50
Obsessive–compulsive disorder (OCD)	F42, (F42.1, F42.2 on SCAN only)	161	8	3	0	−0·03	−0·05 to 0·00
Any of the above anxiety disorders	Any F42, F41, or F40	99	35	13	25	0·33	0·18 to 0·47
Any above ICD-10 depression or anxiety diagnoses							
Diagnoses covered in CIDI and SCAN	Any F32, F33, F34.1, F40, F41, F41.8, F42	96	32	11	33	0·43	0·29 to 0·57

* The proportion of observed agreements adjusted for the rate of agreement that would be expected just by chance.
† Sample size for dysthymia = 170.

Table 2. *Index of agreement, sensitivity and specificity of SCAN and CIDI-Auto ICD*-10 *diagnoses present in the month before interview in second-stage screen positive respondents in the general population* (N = 172)

Diagnosis	Index of agreement*	95% CI	Sensitivity†	95% CI	Specificity‡	95% CI
Depressive disorders						
Any depressive episode	0·14	0·03 to 0·35	0·50	0·12 to 0·88	0·90	0·85 to 0·94
Depressive episode or disorder	0·11	0·02 to 0·28	0·50	0·12 to 0·88	0·87	0·81 to 0·91
Dysthymia§	0·33	0·10 to 0·65	0·57	0·18 to 0·90	0·97	0·93 to 0·99
Any depression above (not remission)	0·29	0·14 to 0·48	0·75	0·43 to 0·95	0·88	0·82 to 0·93
Anxiety disorders						
Any agoraphobic anxiety disorder	0·19	0·04 to 0·46	0·75	0·19 to 0·99	0·93	0·88 to 0·96
Agoraphobia with panic disorder	0·17	0·00 to 0·64	0·50	0·01 to 0·99	0·98	0·94 to 0·99
Social phobias	0·27	0·06 to 0·61	1·00	0·29 to 1·00	0·95	0·91 to 0·98
Specific/isolated phobia	0·29	0·16 to 0·45	0·57	0·34 to 0·78	0·86	0·80 to 0·91
Any phobic anxiety above	0·33	0·20 to 0·47	0·68	0·46 to 0·85	0·82	0·74 to 0·88
Panic disorder	0·19	0·06 to 0·38	0·36	0·13 to 0·65	0·92	0·86 to 0·96
Generalized anxiety disorder (GAD)	0·00	0·00 to 0·17	0·00	0·00 to 0·71	0·90	0·84 to 0·94
Any non-phobic anxiety disorder	0·20	0·09 to 0·36	0·47	0·23 to 0·72	0·85	0·79 to 0·90
Any phobia, panic, or GAD	0·35	0·24 to 0·47	0·67	0·49 to 0·81	0·76	0·68 to 0·83
Obsessive–compulsive disorder (OCD)	0·00	0·00 to 0·28	0·00	0·00 to 0·71	0·95	0·91 to 0·98
Any of the above anxiety disorders	0·34	0·24 to 0·46	0·66	0·49 to 0·80	0·74	0·66 to 0·81
Any above ICD-10 depression or anxiety diagnoses						
Meeting any one or more ICD-10 DCR criteria for above disorders (listed in Table 1)	0·43	0·32 to 0·55	0·75	0·60 to 0·86	0·75	0·68 to 0·82

* Index of agreement: the proportion of subjects for whom a positive diagnosis was recorded by either observer in which both agreed (unlike kappa, Index of Agreement is not sensitive to prevalence).

† Sensitivity: the proportion of true (i.e. SCAN) cases identified by CIDI.

‡ Specificity: the proportion of non-cases (SCAN negative) that were correctly identified as negative by the CIDI.

§ Sample size for dysthymia = 170.

Table 3. *Varying thresholds applied to the total depression scores using ICD*-10 *criteria for depressive disorder* (N = 172)

CIDI total score threshold*	kappa	95% CI kappa
0/1–9	0·12	0·04 to 0·20
0–1/2–9	0·23	0·10 to 0·36
0–2/3–9	0·47	0·31 to 0·62
0–3/4–9	0·47	0·29 to 0·66
0–4/5–9	0·36	0·14 to 0·57
0–5/6–9	0·20	−0·01 to 0·42
0–6/7–9	0·29	0·02 to 0·56
0–7/8–9	0·15	−0·11 to 0·40
ICD-10 depression using common algorithm for both CIDI and SCAN	0·15	−0·03 to 0·34

* CIDI depression score ranges from 0 to 9 yielding eight cut-points representing different severity thresholds (the SCAN total scores ranged from 0 to 8 only).

hypothetical study was of the entire population whereas in the present study, CIDI and SCAN assessments were restricted to those already identified as being at greater risk of developing disorder (using the CIS-R). First, we assumed that 25% of these subjects would be 'at high risk' and undergoing second-stage clinical reappraisals. This was then relaxed to allow the proportion 'at risk' to lie somewhere between 15 and 35%: our prior expectation was that it was equally likely to be anywhere within this range (Table 4).

Rates of disorder differ according to house ownership (Jenkins *et al.* 1997*b*), and diagnostic concordance levels may also differ between such subgroups (or strata). Taking account of this might increase the precision and accuracy of recalibrated estimates (Kessler, 1999). Therefore, as an example, the recalibrated prevalence estimate was recalculated after taking account of home ownership (Table 4).

RESULTS

Two thousand two hundred and fifty addresses were randomly allocated to interviewers. Interviewers had visited 1882 addresses by the time fieldwork was halted. They identified 1170 eligible adults. Ninety-six of the sampled adults refused to take part, and 211 could not be

Table 4. *Corrected estimates calibrated according to SCAN of prevalence of any CIDI diagnosis listed in Table* 1 (*any F*32, *F*33, *F*34.1, *F*40, *F*41, *F*41.8, *F*42) *overall* (N = 172) *and within stratified subgroups by ownership of accommodation* (N = 108) *and according to prior assumptions**

			SCAN/CIDI					
			−/−	−/+	+/−	+/+		
Prior assumptions*	Strata estimate		Both interviews negative	Only CIDI positive	Only SCAN positive	Both interviews positive	Corrected prevalence estimate %	95% CI
25% at high risk	Un-stratified	Overall	96	32	11	33	6·2	4·8 to 7·8
15% to 35% at high risk	Un-stratified	Overall	96	32	11	33	6·2	4·5 to 8·3
	Stratified	Accommodation owned	63	23	6	16	5·1	3·2 to 7·3
		Accommodation rented	33	9	5	17	6·4	3·9 to 9·7
		Overall (combined)	96	32	11	33	5·5	3·8 to 7·5

* Prior assumptions: assumptions we make about the distribution of a variable before new empirical data are considered using Bayesian statistics (conventional statistical methods could produce misleading 95% confidence intervals).

contacted or were ineligible (Brugha *et al.* 1999*b*). Completed screening interviews were returned by 860 individuals: 473 respondents with total scores < 8 on the CIS-R were not interviewed further (Fig. 1); 387 of 860 (45%) who scored ⩾ 8 completed either a SCAN or CIDI diagnostic interview, and 172 of these completed a second diagnostic interview. All interview were carried out within a maximum interval of 15 days and 87 within 5 days.

Demographic characteristics and refusal rates (Fig. 1) were similar to the national survey sample (Meltzer *et al.* 1995; Brugha *et al.* 1999*a*), but a higher proportion was female, not married, members of an ethnic minority and urban dwellers, all factors associated with higher rates of disorder (Jenkins *et al.* 1997*b*). Participation in follow-up interviews was somewhat greater in older respondents, and significantly so in women ($P = 0{\cdot}013$). Other sociodemographic factors, and the presence of depression, anxiety disorder or a high CIS-R score, were not associated. The mean duration of the CIDI interviews was 63 min (range 20 to 185 min) and of the SCAN, 34 min (range 15 to 75).

Concordance findings

Comparisons of CIDI and SCAN ICD-10 study diagnoses yielded coefficients of concordance ranging from a kappa of −0·03 to +0·48. These were in the 'poor' to 'fair' range (Landis & Koch, 1977) except for chronic mild depression (dysthymia), social phobia and any ICD-10 study diagnosis, which were 'moderately' concordant (Table 1). These findings were supported when the statistic 'Index of agreement' was used (Table 2).

Our analysis of co-morbidity produced similar findings. For panic and depression occurring together, kappa was 0·11 (95% CI: −0·14 to 0·36); for any depression diagnosis and any study anxiety disorder, it was 0·23 (95% CI: 0·03 to 0·44); for any two or more disorders (depression, panic, phobia, generalized anxiety disorder, or OCD), it was 0·37 (95% CI: 0·18 to 0·57). Each disorder type classified by SCAN was identified at least once by CIDI. The sensitivity was widely variable, whereas specificity was substantial, ranging from 0·74 to 0·98 (Table 2).

CIDI false positives

Set against the SCAN calibration data there was a consistent pattern of 'false positives' for all CIDI diagnoses. According to the SCAN evaluation, most received mild, below criterion ratings (true positives received 'clinically significant' ratings). Some were positive for the wrong disorder. Only 1 in 7 denied (on SCAN) having any of the criteria of the 'missed' disorder.

Interview order effects

There were interview order effects on concordance but not on prevalence. Although statistically non-significant, when the CIDI preceded the SCAN, concordance appeared somewhat better. For any study ICD-10 disorder, kappa

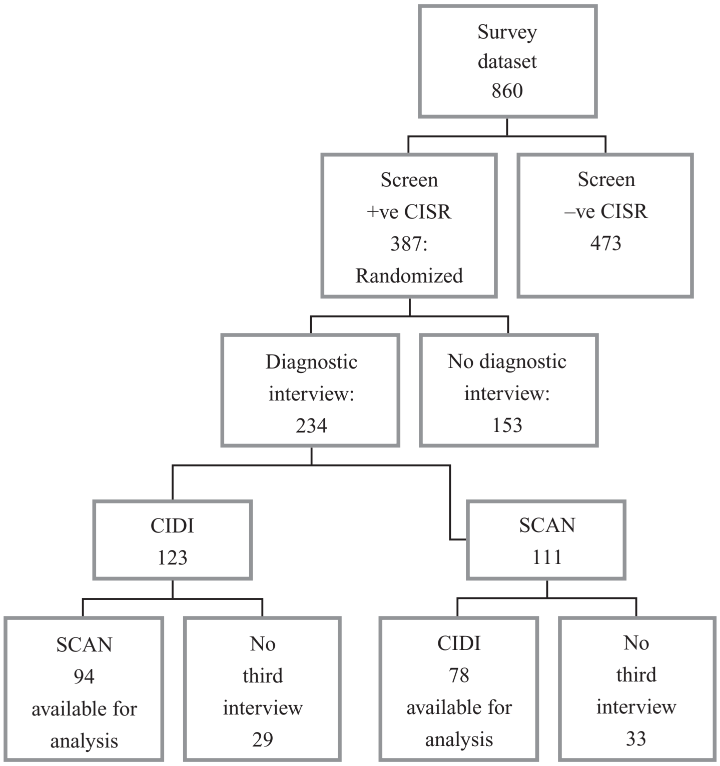

FIG. 1. Sampling and interviews at each stage: yielding $N = 94 + 78$ completed pairs of SCAN and CIDI interviews for analysis (CIDI, structured diagnostic measure administered by lay survey interviewers; SCAN, semi-structured diagnostic assessment administered by clinicians).

was 0·52 (95% confidence interval = 0·34 to 0·70) when the CIDI preceded the SCAN. It was 0·33 (95% confidence interval = 0·12 to 0·54) when it followed SCAN. This trend was particularly marked for anxiety disorders. The time interval length between interviews did not affect concordance.

Sensitivity analyses

The new depression algorithm agreed completely with the published algorithms on the diagnosis of depression (any F32 or F33 disorder). Using the correlation coefficient Kendall's tau for the depression total score (tau = 0·34; $P < 0{\cdot}01$) there was a non-significant improvement on that for the depression diagnostic classification. But when all possible thresholds were evaluated using each level on the depression criteria score (Table 3) there was more pronounced variation in concordance: kappa ranged from approximately 0·5 to 0·2 for cut-points at the middle and extremes of the scale respectively (Table 3).

Sensitivity analyses confirmed that the lack of second stage information on persons only assessed in the first stage did not significantly bias the principal concordance findings. When we adjusted for putative screen negatives (Brugha *et al.* 1999*b*) by adding them to the cell concordant for 'non-cases', the coefficient for any ICD-10 study disorder changed from 0·43 to 0·46.

For any ICD-10 study disorder, the proportion positive on CIDI was 41% and 35% in Group A and B respectively. Concordance (kappa) differed between interviewer groups: 0·51 (95% CI; 0·32, 0·69) and 0·35 (95% CI; 0·15, 0·55) respectively. Concordance for any depressive disorder was 0·08 and 0·24 respectively, while Kendall's tau for the depression total score was identical for the two interviewers (tau = 0·34; $P < 0{\cdot}01$).

Recalibrated prevalence estimates

The recalibration exercise reduced the prevalence for any study depressive or anxiety disorder from 9·0 to 6·2 % (Table 4). When we relaxed the assumption to allow the proportion completing a SCAN interview to lie somewhere between 15 and 35 %, the resulting prevalence estimate was also 6·2 % but, as one would expect, precision was less (Table 4).

Given the result of their CIDI assessment, people were more likely to be diagnosed positive with the SCAN if they were living in rented accommodation (Poisson modelling, $P = 0{\cdot}044$) (Table 3). Taking account of this reduced the overall prevalence estimate to 5·5 % (Table 4), slightly lower than the previous estimate of 6·2 %, but estimated with almost equal precision.

The separate recalibrated prevalence estimates for the two SCAN interviewers were 6·8 % (95 % CI: 4·9 to 9·6 %) and 5·4 % (95 % CI: 3·3 to 8·1 %) respectively. The proportion of CIDI negatives that were positive on SCAN was equal in both groups (10 %), but the proportion of CIDI positives that were positive on SCAN was small and not significantly different (59 and 42 % respectively).

DISCUSSION

This report compares the Lifetime CIDI with reference to SCAN as a method of case identification. It is the first study of its kind specifically designed for the purpose. Its advantages over previous comparison studies include the use of a separate random sample of the general population (Brugha *et al.* 1999*a*), the random ordering of instruments, and blinded assessment in order to minimize rater bias (Goodie & Fantino, 1996). To maximize co-operation, we restricted our study to anxiety and depressive disorders present during the month before interview.

We found 'poor' to 'fair' concordance across all specific study disorders, except for dysthymia and social phobia, for which concordance was 'moderate' (Landis & Koch, 1977). Concordance for co-morbidity (not previously evaluated in such a study) was also poor to fair. An examination of order effects suggested that comparisons in earlier surveys (Brugha *et al.* 1999*b*) between fully structured and reference measures could have produced moderately biased over-estimates of concordance.

Concordance for current depressive episode or disorder was 'low' (Landis & Koch, 1977) (kappa = 0·15), whether using published algorithms or a specially developed new algorithm common to both interviews. As expected, the depression score was more concordant than a diagnostic ('binary') classification. Lowering the defined threshold for depression improved concordance (Table 3), although it still lay outside the acceptable range. However, a very low threshold (i.e. at least one, or $\geqslant 2$ depression criteria = a case of depression) yielded poor concordance.

When corresponding analyses were carried out in a comparison of the fully structured CIS-R and SCAN (Brugha *et al.* 1999*b*), similar but less pronounced variations in concordance emerged. Further replications are needed before firm conclusions are drawn.

Possible study limitations

There were constraints on the representativeness of the sample. More persons were at risk of disorder than in our national survey (Jenkins *et al.* 1997*b*), as reflected in the unadjusted proportions scoring at or above an extremely sensitive subthreshold level of 8 on the CIS-R (45 %) in the present study compared with the national survey in Great Britain (28 %). Many respondents refused to take part in two diagnostic interviews in addition to the initial screening interview (Fig. 1). However, only male gender significantly predicted non-participation. We compared SCAN and screening interview diagnostic agreement levels (Brugha *et al.* 1999*b*) between 33 respondents who refused the third interview and 78 who completed it (Fig. 1), but there was no effect of refusal on concordance. Neither was concordance related to gender.

Our concordance estimates are based on the evidence-based assumption that screen negatives do not fulfil criteria for any diagnosis (Brugha *et al.* 1999*b*). Sensitivity analyses whereby false negatives are treated as concordant cases gave only marginally different results, but still showed moderate concordance. Our use of Bayesian statistics allowed us to take into account prior assumptions about respondents who did not undergo a SCAN interview at the second stage.

Some authorities, although agreeing that the inflexible approach to questioning used in fully structured interviews can lead to an increased risk of invalidity with regard to some diagnoses, have strong reservations about the nature of the appropriate reference measure for a fully structured diagnostic instrument like CIDI (Wittchen *et al.* 1999). We argue on theoretical grounds that semi-structured clinical interviews should provisionally be accepted as the reference standard because their face validity is apparent in their similarity to assessment procedures in clinical psychiatric practice (Brugha *et al.* 1999*a*). Although our position is not fully accepted (Wittchen *et al.* 1999), this approach to the validation of structured interviews is widely employed (Wittchen, 1994; Kessler, 1999). Fully structured interviews would have greater claim to gold standard status if their external validity turned out to be superior. This would however radically alter what is meant by diagnosis (Spitzer, 1983). Since collateral psychiatric information is unavailable for the vast majority of community survey respondents, comparisons with best consensus or 'LEAD' diagnoses does not offer a feasible reference point.

Standardization may be harder to achieve when clinical judgement is required of interviewers because of increased interviewer variation (Sturt *et al.* 1981; Wittchen *et al.* 1999). Further analysis in the current study did reveal some differences between our SCAN interviewers, but the principle findings were not altered. Although the CIDI-based prevalence of any ICD-10 study disorder varied between the respondents of our two SCAN interviewers, we cannot tell if this is due to differences in case prevalence (Cicchetti & Feinstein, 1990), in rating thresholds (Sturt *et al.* 1981), or in how respondents reveal information about their symptoms to different interviewers (Reissman, 1977).

There is also a problem with the reproducibility of diagnoses generated using semi-structured 'clinical' interviews in the general population. Although good 'present-state' test-retest reliability has been shown in clinical populations with SCAN (Brugha *et al.* 1999*c*) studies of reproducibility in the general population are clearly a priority.

Interpretation of findings

Prevalence of study anxiety or depressive disorder in the past month was lower with SCAN than with CIDI. This finding contrasts with recent comparisons of lifetime prevalence obtained from semi-structured and fully structured community diagnostic interviews (Regier, 2000). As our analyses of 'CIDI false positives' showed, this appears to reflect the stricter requirement in SCAN that each symptom must be distressing, difficult to control and excessive under the circumstances in order for it to be rated (World Health Organization Division of Mental Health, 1992). Thus, many endorsed CIDI items were clearly considered by the SCAN interviewer but judged as subthreshold. Another explanation is that this severity difference might be attributed to a mode effect such as concealment of embarrassing information from clinical interviewers (Turner *et al.* 1988). Concurrent comparisons of audiocomputer and clinical interviewing in primary care have indeed demonstrated marked under-reporting of socially sensitive behaviours in the latter, i.e. HIV risk behaviour and alcohol misuse (Kobak *et al.* 1997; Turner *et al.* 1998), but under-reporting of anxiety and major depression did not occur (Kobak *et al.* 1997). Mean Hamilton Rating Scale depression (Kobak, 1996) scores rated by clinicians were somewhat lower than computerized assessments (Mundt *et al.* 1998). This did not apply to anxiety (Kobak, 1996). In contrast, a long-term follow-up (median, 12·6 years) of the Baltimore ECA cohort revealed greater reporting of depression with SCAN compared to the fully structured DIS (Eaton *et al.* 2000). The median delay between interviews in this study was 113 days. Thus, consistent mode effect findings for anxiety and depression are not seen in the literature, and therefore do not bear on our conclusions. However, the time frame and content of assessments appear important.

Could the CIDI be particularly limited when recording disorders during the past month, as reported for the DIS (Anthony *et al.* 1985)? Only the SCAN interview requires a check that each criterion is met currently. Evaluating the validity of past year or lifetime ever CIDI diagnoses might iron out short-term fluctuations in mental state, generating higher concordances. However, recall problems (Kessler *et al.* 1998; Andrews *et al.* 1999) might substantially undermine the validity of both retrospective assessments. In a comparison of the lay administered DIS (Helzer *et al.* 1985) with criterion checklists applied by physicians, agreement for lifetime ever DSM-III depressive and anxiety disorders was also fair in general population respondents.

Previous studies of the validity of fully structured diagnostic interviews may well have over-estimated concordance. Our tentative finding of interview order effects needs to be independently replicated. We expected semi-structured interviewing to have a greater educational effect on respondents than fully structured interviewing. This would enhance concordance when the clinical interview comes first. However, we found the opposite.

The suggestion that low concordance in community studies is attributable to mild and relatively infrequent cases (Helzer *et al.* 1985) is partly justified by our analyses. They confirm the need to interpret kappa cautiously for very rare occurrences (Cicchetti & Feinstein, 1990), as with the more specific disorder categories in Table 1. Information on morbidity is more likely to be valid if full scores or mid-scale points are used instead of rigidly dichotomized categories (Brugha *et al.* 1999*b*; Eaton *et al.* 2000).

A way forward?

It should not be so surprising that different modes of interviewing yield such different information, and it may be that they can be used in complementary ways.

Using a method based on Bayes' Theorem, we calculated a prevalence estimate, with 95% confidence intervals 'as if' SCAN had been used in a large scale survey as a first stage interview. Surprisingly, the poor concordance between CIDI and SCAN (Table 1) did not make it impossible to recalibrate the prevalence estimate in this way. The overall accuracy of the recalibrated prevalence estimate was not altered by the addition of a stratified analysis (Kessler, 1999) based on home ownership, although this covariate is significantly associated with prevalence (Jenkins *et al.* 1997*b*). Precision was also limited by the statistical power of the present sample size. Differences between the two SCAN interviewers in the level of 'correction' applied were not significant.

Once population and sampling differences have been adjusted, the method for correcting for prevalence differences could be used in other large-scale general population survey data sets using CIDI. The correct classification of individual respondents in terms of potential need of health care would improve estimates of population need for treatment and the services that deliver it (Brewin *et al.* 1987). Methods for imputing for missing data (Little & Rubin, 1987) on respondents not completing a clinical reappraisal have been proposed for this purpose (Kessler, 1999). The incorporation of clinical reappraisals into recent large-scale surveys (Singleton *et al.* 1998; Meltzer *et al.* 2000; Office for National Statistics Social Survey Division, 2000) and those now underway (Kessler, 1999) could thus greatly enhance the information value to be gained from them.

We have discussed elsewhere (Brugha *et al.* 1999*b*) the steps that might be taken to minimize measurement error, due for example to differences between interviewers at the training, data collection and data analysis stages (van der Zouwen *et al.* 1991). Diagnostic validity could also be enhanced using other strategies (Brugha *et al.* 1999*a*), for example, greater incorporation of clinical judgement into lay survey interview methods (Brugha *et al.* 1999*a*, *d*). Efforts to improve concordance between fully structured and semi-structured measures (Kessler *et al.* 1988; Kessler, 1999) and the reliability of both approaches should be continued.

Until survey data using clinical interviewer-based interviewing strategies that use clinical judgment to rate symptoms are available and evaluated (Kessler, 1999), epidemiologists and public health clinicians should interpret the findings of existing surveys in the general population in the light of this and earlier comparison studies (Anthony *et al.* 1985; Helzer *et al.* 1985; Brugha *et al.* 1999*a*, *b*).

The Statistics Division, Department of Health, London, provided grant funding. Professor Anne Farmer assisted with CIDI training. The field work interviewers were: Dr Jo Vernon and Dr Mark Janas, Ursula Bilson, Janice Briggs, Anne Dillon, Shavnam Dosanjh, Annette Garner, Sheila Goward, Rebecca Legard, Ron Moreton, Dianna Oxley, Tony Pritchett, Jane Smith and Suzanne Window. We are grateful to C. Hendricks Brown for his encouragement and suggestions regarding recalibration methods.

REFERENCES

Andrews, G., Anstey, K., Brodaty, H., Issakidis, C. & Luscombe, G. (1999). Recall of depressive episode 25 years previously. *Psychological Medicine* **29**, 787–791.

Anthony, J. C., Folstein, M. F., Romanoski, A. J., Von Korff, M., Nestadt, G. R., Chahal, R., Merchant, A., Brown, C. H., Shapiro, S., Kramer, M. & Gruenberg, E. M. (1985). Comparison of Lay Diagnostic Interview Schedule and a standardised psychiatric diagnosis. *Archives of General Psychiatry* **42**, 667–675.

Bartlett, C. J. & Coles, E. C. (1988). Psychological health and well-being: why and how should public health specialists measure it? Part 1: rationale and methods of the investigation, and review of psychiatric epidemiology. *Journal of Public Health Medicine* **20**, 281–287.

Biemer, P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. & Sudman, S. (1991). *Measurement Errors in Surveys*. John Wiley & Sons: New York.

Brewin, C. R., Wing, J. K., Mangen, S. P., Brugha, T. S. & MacCarthy, B. (1987). Principles and practices of measuring needs in the long-term mentally ill: the MRC needs for care assessment. *Psychological Medicine* **17**, 971–981.

Bromet, E. J., Dunn, L. O., Connell, M. M., Dew, M. A. & Schulberg, H. C. (1986). Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry* **43**, 435–440.

Brugha, T. S., Jenkins, R., Bebbington, P., Meltzer, H. & Taub, N. A. (1997). The scope for increasing the usefulness of population based epidemiological information on need. University of New South Wales, Clinical Research Unit for Anxiety Disorders: Sydney, NSW, Australia.

Brugha, T. S., Bebbington, P. E. & Jenkins, R. (1999*a*). A difference that matters: comparisons of structured and semi-structured diagnostic interviews of adults in the general population. *Psychological Medicine* **29**, 1013–1020.

Brugha, T. S., Bebbington, P., Jenkins, R., Meltzer, H., Taub, N. A., Janas, M. & Vernon, J. (1999*b*). Cross validation of a household population survey diagnostic interview: a comparison of CIS-R with SCAN ICD-10 diagnostic categories. *Psychological Medicine* **29**, 1029–1042.

Brugha, T. S., Nienhuis, F. J., Bagchi, D., Smith, J. & Meltzer, H. (1999*c*). The survey form of SCAN: the feasibility of using experienced lay survey interviewers to administer a semi-structured systematic clinical assessment of psychotic and non psychotic disorders. *Psychological Medicine* **29**, 703–712.

Brugha, T. S., Taub, N. A., Bebbington, P. E., Jenkins, R. & Meltzer, H. (1999*d*). The validity of an international structured diagnostic interview (CIDI 1.1) in the general population. Poster presentation to the International Epidemiology Association Biennial Congress (IEA99), Florence, Italy.

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

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Der, G., Glover, G., Brugha, T. S. & Wing, J. K. (1998). SCAN version 1: algorithms and CAPSE 10.1. In *Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN/PSE-10* (ed. J. K. Wing, N. Sartorius and T. B. Üstün), pp. 110–115. Cambridge University Press: Cambridge.

Eaton, W. W., Neufeld, K., Chen, L. S. & Cai, G. (2000). A comparison of self-report and clinical diagnostic interviews for depression: diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the Baltimore epidemiologic catchment area follow-up. *Archives of General Psychiatry* **57**, 217–222.

Everitt, B. S. & Der, G. (1996). *A Handbook of Statistical Analysis Using SAS*. Chapman and Hall: London.

Goodie, A. S. & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature* **380**, 247–249.

Helzer, J. E., Robins, L. N., McEvoy, L. T., Spitznagel, E. L., Stoltzman, R. K., Farmer, A. & Brockington, I. 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.

Jenkins, R., Bebbington, P., Brugha, T., Farrell, M., Gill, B., Lewis, G., Meltzer, H. & Petticrew, M. (1997*a*). The national psychiatric morbidity surveys of Great Britain – strategy and methods. *Psychological Medicine* **27**, 765–774.

Jenkins, R., Lewis, G., Bebbington, P., Brugha, T., Farrell, M., Gill, B. & Meltzer, H. (1997*b*). The national psychiatric morbidity surveys of Great Britain – initial findings from the household survey. *Psychological Medicine* **27**, 775–789.

Kessler, R. C. (1999). The World Health Organization International Consortium in Psychiatric Epidemiology (ICPE): initial work and future directions – the NAPE Lecture 1998. Nordic Association for Psychiatric Epidemiology. *Acta Psychiatrica Scandinavica* **99**, 2–9.

Kessler, R. C., Wittchen, H. U., Abelson, J. M., McGonagle, K. A., Schwarz, N., Kendler, K. S., Knauper, B. & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS). *International Journal of Methods in Psychiatric Research* **7**, 33–55.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons Ltd: London.

Kobak, K. A. (1996). Clinical computing. Computer-administered symptom rating scales. *Psychiatric Services* **47**, 367–369.

Kobak, K. A., Taylor, L. H., Dottl, S. L., Greist, J. H., Jefferson, J. W., Burroughs, D., Mantle, J. M., Katzelnick, D. J., Norton, R., Henk, H. J. & Serlin, R. C. (1997). A computer-administered telephone interview to identify mental disorders. *Journal of the American Medical Association* **278**, 905–910.

Kruskal, W. (1991). Introduction. In *Measurement Errors in Surveys* (ed. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman), pp. xxiii–xxxiii. John Wiley & Sons: New York.

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

Leeman, E. (1998). Misuse of psychiatric epidemiology. *Lancet* **351**, 1601–1602.

Lewis, G., Pelosi, A. J., Araya, R. & Dunn, G. (1992). Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine* **22**, 465–486.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. John Wiley: New York.

McLeod, J. D., Turnbull, J. E., Kessler, R. C. & Abelson, J. M. (1990). Sources of discrepancy in the comparison of a lay-administered diagnostic instrument with clinical diagnosis. *Psychiatry Research* **31**, 145–159.

Marcus, S. C. & Robins, L. N. (1998). Detecting errors in a scoring program: a method of double diagnosis using a computer-generated sample. *Social Psychiatry and Psychiatric Epidemiology* **33**, 258–262.

Meltzer, H., Gill, B., Petticrew, M. & Hinds, K. (eds.) (1995). *Office of Population Censuses & Surveys Social Survey Division. OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 1: The Prevalence of Psychiatric Morbidity among Adults Living in Private Households*. OPCS Surveys of Psychiatric Morbidity in Great Britain. Her Majesty's Stationery Office: London.

Meltzer, H., Gatward, R., Goodman, R. & Ford, T. (eds.) (2000). *Office for National Statistics Social Survey Division. Mental Health of Children and Adolescents in Great Britian*. The Stationery Office. OPCS Surveys of Psychiatric Morbidity in Great Britain: London.

Mundt, J. C., Kobak, K. A., Taylor, L. V., Mantle, J. M., Jefferson, J. W., Katzelnick, D. J. & Greist, J. H. (1998). Administration of the Hamilton Depression Rating Scale using interactive voice response technology. *MD Computing* **15**, 31–39.

Murray, C. J. & Lopez, A. D. (1997). Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet* **349**, 1436–1442.

Office for National Statistics Social Survey Division (2000). *Survey of Psychiatric Morbidity Among Adults in Great Britain, 2000*. Office for National Statistics: London.

Pickles, A., Dunn, G. & Vázquez-Barquero, J. L. (1995). Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research* **4**, 73–89.

Regier, D. A. (2000). Community diagnosis counts. *Archives of General Psychiatry* **57**, 223–224.

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C. & Norquist, G. S. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders. *Archives of General Psychiatry* **55**, 109–115.

Reissman, C. K. (1977). Interviewer effects in psychiatric epidemiology: a study of medical and lay interviewers and their impact on reported symptoms. *American Journal of Public Health* **69**, 485–491.

Robins, L. 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, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., Farmer, A., Jablenski, A., Pickens, R., Regier, D. A., Sartorius, N. & Towle, M. S. (1988). The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

Romanoski, A. J., Nestadt, G., Cahal, R., Merchant, A., Folstein, M. F., Gruenberg, E. M. & McHugh, P. R. (1988). Interobserver reliability of a 'Standard Psychiatric Examination' (SPE) for case ascertainment (DSM-III). *Journal of Nervous and Mental Disease* **176**, 63–71.

Singleton, N., Meltzer, H., Gatward, R., Coid, J. & Deasy, D. (eds.) (1998). *Office for National Statistics Social Survey Division. Psychiatric Morbidity among Prisoners in England and Wales. A survey carried out in 1997 by the Social Survey Division of ONS on behalf of the Department of Health*. The Stationery Office. ONS Surveys of Psychiatric Morbidity in Great Britain: London.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*. MRC Biostatistics Unit: London.

Spitzer, R. L. (1983). Psychiatric diagnosis: are clinicians still necessary? *Comprehensive Psychiatry* **24**, 399–411.

Spitzer, R. L., Williams, J. B., Gibbon, M. & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. *Archives of General Psychiatry* **49**, 624–629.

Sturt, E., Bebbington, P. E., Hurry, J. & Tennant, C. (1981). The Present State Examination used by interviewers from a survey agency: Report from the Camberwell Community Survey. *Psychological Medicine* **11**, 185–192.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H. & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* **280**, 867–873.

Üstün, T. B., Harrison, G. L. & Chatterji, S. (1998). Training in the use of SCAN. In *Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN/PSE-10* (ed. J. K. Wing, N. Sartorius and T. B. Üstün), pp. 68–85. Cambridge University Press: Cambridge.

van der Zouwen, J., Dijkstra, W. & Smit, H. (1991). Studying respondent-interviewer interaction: the relationship between interviewing style, interviewer behaviour, and response behaviour. In *Measurement Errors in Surveys* (ed. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman), pp. 419–461. John Wiley & Sons: New York.

Wilson, P. R. & Elliot, D. J. (1987). An evaluation of the Postcode Address File as a sampling frame and its use within OPCS. *Journal of the Royal Statistical Society Series A* (*General*) **150**, 230–240.

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D. & Sartorius, N. (1990). SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* **47**, 589–593.

Wing, J. K., Sartorius, N. & Üstün, T. B. (1998). Measurement and classification in psychiatry. In *Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN/PSE-10* (ed. J. K. Wing, N. Sartorius & T. B. Üstün), pp. 1–11. Cambridge University Press: Cambridge.

Wittchen, H. U. (1994). Reliability and validity studies of the WHO – Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.

Wittchen, H. U., Üstün, B. & Kessler, R. C. (1999). Diagnosing mental disorders in the community: a difference that matters? *Psychological Medicine* **29**, 1021–1027.

World Health Organization Division of Mental Health (1992). *SCAN Schedules for Clinical Assessment in Neuropsychiatry, Version 1.0*. World Health Organization: Geneva.

World Health Organization (1993*a*). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. WHO: Geneva.

World Health Organization (1993*b*). *CIDI-Auto Version 1.1: Administrator's Guide and Reference. 1.1d* (*Release 2.04, January 1994*). 0-646-16372-8. MS-DOS. World Health Organization: Geneva.