Psychological Medicine, 1999, **29**, 1029–1042. Printed in the United Kingdom

Cross validation of a general population survey diagnostic interview: a comparison of CIS-R with SCAN ICD-10 diagnostic categories

T. S. BRUGHA,[1] P. E. BEBBINGTON, R. JENKINS, H. MELTZER, N. A. TAUB, M. JANAS AND J. VERNON

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

ABSTRACT

Background. Comparisons of structured diagnostic interviews with clinical assessments in general population samples show marked discrepancies. In order to validate the CIS-R, a fully structured diagnostic interview used for the National Survey of Psychiatric Morbidity in Great Britain, it was compared with SCAN, a standard, semi-structured, clinical assessment.

Methods. A random sample of 1882 Leicestershire addresses from the Postcode Address File yielded 1157 eligible adults: of these 860 completed the CIS-R; 387 adults scores $\geqslant 8$ on the CIS-R and 205 of these completed a SCAN reference examination. Neurotic symptoms, in the previous week and month only, were enquired about. Concordance was estimated for ICD-10 neurotic and depressive disorders, F32 to F42 and for depression symptom score.

Results. Sociodemographic characteristics closely resembled National Survey and 1991 census profiles. Concordance was poor for any ICD-10 neurotic disorder (kappa = 0·25 (95% CI, 0·1–0·4)) and for depressive disorder (kappa = 0·23 (95% CI, 0–0·46)). Sensitivity to the SCAN reference classification was also poor. Specificity ranged from 0·8 to 0·9. Rank order correlation for total depression symptoms was 0·43 (Kendall's tau b; $P < 0{\cdot}001$; $N = 205$).

Discussion. High specificity indicates that the CIS-R and SCAN agree that prevalence rates for specific disorders are low compared with estimates in some community surveys. We have revealed substantial discrepancies in case finding. Therefore, published data on service utilization designed to estimate unmet need in populations requires re-interpretation. The value of large-scale CIS-R survey data can be enhanced considerably by the incorporation of concurrent semi-structured clinical assessments.

INTRODUCTION

There are, broadly, two types of assessment of mental disorders that can be used in epidemiological surveys: semi-structured clinical interviews and lay-administered structured diagnostic questionnaires. Each has a characteristic approach to cost-effectiveness and each has advantages and disadvantages (Brugha *et al.* 1999*a*). In ordinary clinical assessment, the experience of the clinician is used as a mechanism for arriving rapidly at a position in which a decision can be made over whether the respondent's experience matches the symptom concept, such that the symptom can be rated. The approach taken with semi-structured instruments such as the Present State Examination (PSE) (Wing *et al.* 1974) approximates most closely to the standards of clinical assessment (Brugha *et al.* 1999*a*). The cost of relying on

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

clinician judgement may be a loss of control over standardization and therefore a potential reduction in reliability (Lewis *et al.* 1992). Reliability is nevertheless maximized by standardizing the items to be covered, providing a structure (with rules about cut-off procedures), and training interviewers to criteria embodied in a common standard defined in a glossary (Wing *et al.* 1974).

The lay questionnaire seeks to reduce cost by eliminating the need for an experienced clinician, but at the expense of validity, since there is no clinical evaluation of responses before a symptom record is made. It is conceivable that the need for clinical judgement in this process could be replaced by considering all possible responses and framing follow-up questions to deal with them. However, this could become an exhaustive, exhausting and inherently unreliable procedure (Brugha *et al.* 1996). For questionnaires to be restricted to an acceptable and feasible length some loss of validity will follow. These trade-offs between feasibility, cost, reliability and validity are inevitable and apparent in every instrument. In this report we focus mainly on the validity of instruments, although we will also comment on feasibility and cost.

In clinical populations, both methods have shown good test–retest reliability, i.e. when re-administered by an independent interviewer after an interval (Wing *et al.* 1998). Validity testing implies an asymmetrical relationship between two instruments, such that one is regarded *a priori* as a gold standard. Most investigators agree that a systematic clinical assessment is the standard by which to assess lay measures (Spitzer, 1983; Helzer *et al.* 1985; Wittchen, 1994; Brugha *et al.* 1999*a*).

Once symptom ratings and responses to questions in structured questionnaires have been recorded, standardized diagnostic classification rules (World Health Organization, 1993) can be applied by computer algorithms. By using identical classification rules in both kinds of interview, direct comparisons of the diagnostic output should reveal only differences in the method of interview assessment (Brugha *et al.* 1999*a*).

Comparison studies have been conducted before, in a variety of clinical settings, using structured interviews such as the Diagnostic Interview Schedule (DIS) (Robins *et al.* 1981) and the Composite International Diagnostic Interview (CIDI) (Robins *et al.* 1988). The instruments were usually used in a rigidly structured format (Spengler & Wittchen, 1988). In some cases a very limited element of clinical flexibility required extended interviewer training (Robins *et al.* 1981). These have been compared with clinical measures such as a modified version of the PSE or clinician diagnoses (McLeod *et al.* 1990; Farmer *et al.* 1991; Wittchen *et al.* 1991; Janca *et al.* 1992; Kovess *et al.* 1992; Wittchen, 1994; Andrews *et al.* 1995). In general, good to excellent coefficients of concordance have been found for most diagnostic sections in clinical populations (Wittchen, 1994).

However, validation studies should be carried out where the measures are intended to be used (Wittchen, 1994), i.e. in the general population. Respondents in such settings may differ from clinical samples, for example, in their unfamiliarity with psychiatric terms such as anxiety, phobia, panic and obsession, by understanding them less precisely (Brugha *et al.* 1999*a*). In clinical settings, symptoms are also more severe. Interviewers may also differ, for example, in their expectation of symptoms and of disorders in the respondent that may influence judgement, giving rise to base rate errors (Goodie & Fantino, 1996). They may communicate their expectations to respondents. Good agreement with clinical examination is less guaranteed in the general population, where disorders are relatively less common.

In samples drawn from the general population, and not selected for particular diagnoses (Brugha *et al.* 1999*a*), levels of concordance have been in the poor range (Anthony *et al.* 1985; Helzer *et al.* 1985; McLeod *et al.* 1990). The forerunner of the CIDI (the earlier Diagnostic Interview Schedule, or DIS) was evaluated on 802 subjects in a general population sample in Baltimore (Anthony *et al.* 1985). The DIS was compared with a clinical measure, based on the PSE, and conducted by trained psychiatrists. The DIS performed very poorly as a measure of prevalence of symptoms and common diagnoses in the month prior to interview, revealing poor agreement on which individuals were cases of specified functional psychiatric disorders.

Helzer and his colleagues (1985) compared DIS interviews administered by clinicians and lay interviewers with DSM-III criterion checklists applied by clinicians in a sample of general population respondents. Agreement for lifetime-ever diagnosis of depressive and neurotic disorders was also poor.

In another study, a comparison of the DIS and the SADS-L, a semi-structured interview administered by clinicians (in this study psychiatric social workers) was carried out in Chicago householders (Spitzer *et al.* 1978; McLeod *et al.* 1990). This revealed substantial discrepancies in the RDC diagnosis of depression over a period of 6 months, which was attributed to recall error, and particularly to inconsistent reports of episode timing within the 6 month period covered by both measures.

Particular problems besetting these studies might have increased discrepancies between assessments: the time elapsing between interviews and the difficulty in specifying the time period to be covered by the interviewers (Helzer *et al.* 1985; McLeod *et al.* 1990).

It is also important to ascertain whether bias exists, leading to a survey measure significantly over-estimating or under-estimating prevalence and to assess sensitivity and specificity in relation to the chosen reference measure. There are increasing concerns about widely varying estimates of prevalence between large scale surveys (Leeman, 1998; Regier *et al.* 1998).

The national survey of psychiatric morbidity comparison study

The private household survey of the National Surveys of Psychiatric Morbidity in Great Britain (Jenkins *et al.* 1997*a*, *b*) employed a fully structured interview to assess neurosis, the Clinical Interview Schedule Revised (CIS-R) (Lewis *et al.* 1992). This was administered by survey interviewers from the Office of Population Censuses and Surveys (OPCS; now the Office for National Statistics (ONS)). The CIS-R was developed from the Clinical Interview Schedule (CIS) (Goldberg *et al.* 1970), which was designed to be used by clinically experienced interviewers, such as psychiatrists, and included ratings requiring clinical judgement. The CIS-R was developed as a fully structured interview, to make it possible for lay survey interviewers to gather information about common (neurotic) psychiatric symptoms during health survey interviews. The CIS-R makes use of questions about the present (i.e. past month and week), with the aim of enhancing recall accuracy. Diagnostic algorithms for selected neurotic and depressive disorders in a range from ICD-10 F32 to F42 were also developed. Extra questions on panic, phobia and autonomic symptoms of anxiety, were added to the CIS-R, which are not in the CIS (Lewis *et al.* 1992). Prevalence estimates based on the CIS-R take account of time periods specified in ICD-10, for example the 2-week rule for depression. Impairment of functioning is covered in the final section of the CIS-R and was used in the depression algorithm. Two small evaluations in clinic attenders showed good agreement between lay and clinician CIS-R interviewers on the overall current severity of illness (Lewis *et al.* 1992).

Aim and objectives

This paper reports on a validation of the CIS-R in a general population sample (Brugha *et al.* 1997*b*). We compared it with a clinically based interview, the enhanced 10th edition of the PSE, that forms the core of the larger instrument, the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing *et al.* 1990; World Health Organization Division of Mental Health, 1992), chosen as the reference or standard measure. We report elsewhere a parallel comparison of the SCAN and the CIDI (Brugha *et al.* 1997*b*), and in preparation (Brugha *et al.* 1999*b*). To examine whether agreement was affected by the imposition of categories on dimensional data, recently termed 'carving nature at its joints' (Kendler & Gardner, 1998), we made comparisons of symptom scores as well as at the diagnostic level.

METHOD

Design

A two-phase survey in the adult general population in urban, suburban and rural areas of Leicestershire was carried out. Second phase sample respondents were selected in order to include all possible diagnosable cases of psychiatric disorder. All were asked to undergo two further interviews, the SCAN and CIDI. We compared the CIS-R and SCAN for diagnoses and for depression severity. Power calculations for kappa (Cohen, 1968) showed that at least 150 successful pairs of interviews would provide

80% power to detect a non-zero kappa at the 5% significance level, if the true population value of kappa was 0·4, given at least a 4% 'prevalence' of disorder in the second phase sample.

Sampling

To minimize costs and maximize successful completion of second phase interviews, the research team worked in a single geographic area in urban, surburban and rural postcode areas in Leicestershire. They were chosen to have socio-economic characteristics representative of Great Britain as a whole. The small users Postcode Address File (PAF) was used as a sampling frame (Wilson & Elliott, 1987). Addresses were drawn randomly by computer from the PAF by the sampling division of OPCS. Addresses were also randomly allocated to interviewers to minimize bias due to expectations about the mental health of people living in particular areas or due to convenience factors. Occupied properties were identified when interviewers visited the chosen addresses during fieldwork throughout 1995. Of 2251 addresses allocated to interviewers, 1873 were visited by December 1995 and no biases were found in checks on addresses completed (Brugha *et al.* 1997*b*).

Subjects

As in the national private household survey (Jenkins *et al.* 1997*a*), (Meltzer *et al.* 1995*a*) eligible adults had to be living at a private household address, as defined by OPCS, and aged 16 to 64 years. Adults currently resident elsewhere, for example in a large long-term residential institution, such as a large student residence or armed forces facility, were not included. One adult in each household was eligible for selection. In the case of multiple adult households random sampling was used (Kish, 1965).

Measures and procedures

Two clinicians and 12 lay survey interviewers were recruited and trained in the sampling techniques used by OPCS in the National Surveys (Meltzer *et al.* 1995*a*) and in structured interviewing with the CIS-R and the CIDI-Auto.

CIS-R

As in the National Survey of Psychiatric Morbidity (Meltzer *et al.* 1995*a*) CIS-R interviewers in this comparison study did not require a clinical background. Training was provided initially by senior trainers from the ONS and by the first author and the course was recorded on videotape. Further courses with additional interviewers made use of these and were conducted by existing interviewers who were behavioural science graduates. Clinical terms were explained by experienced clinicians who had also undergone training in the CIS-R and the CIDI. An experienced survey interviewer supervised them.

SIFT

Using national survey data, 'SIFT positive' respondents were predicted to yield prevalence rates at least four times greater than that of the base population for the disorders under study, by using a cut-off $\geqslant 8$ on the CIS-R total score. The number of cases with an ICD-10 diagnosis of psychiatric disorder below the threshold (< 8) was also determined from national CIS-R data (H.M.). SCAN cases would also be expected to be extremely infrequent. SIFT positive respondents were asked to complete an assessment, as soon as possible, with one of the two other survey instruments (SCAN or CIDI) within 14 days, followed by the remaining instrument, within another 14 days.

SCAN

Two clinicians, with at least 3 years experience in full-time psychiatric practice, were trained in the SCAN (Wing *et al.* 1990). Before they were permitted to undertake unsupervised pilot and survey interviews, senior trainers (T.S.B., P.E.B.) who rated interviews alongside them established the reliability of their ratings. All sections of SCAN Version 1 (World Health Organization Division of Mental Health, 1992) were covered apart from the drug and alcohol sections.

Rating periods

The CIS-R and SCAN interviews make use of different rating periods. In order to permit comparisons with the CIS-R, ratings for the week preceding the CIS-R interview were coded separately alongside the usual Present Episode (PE) or Present State (PS) ratings (Wing *et al.* 1990). The CIS-R was always conducted first. The SCAN interviewers were given the date of the CIS-R interview in order to record ratings for the week preceding it.

Symptoms and diagnostic classifications

CIS-R ICD-10 disorders were derived from the algorithm, based on the published Diagnostic Criteria for Research (World Health Organization, 1993), used in the National Survey of Psychiatric Morbidity (Meltzer *et al.* 1995*a*). The official ICD-10 algorithm for SCAN (version 29 March, 1994) (Der & Wing, 1992), developed under the auspices of WHO, was based on the field trial version of the Diagnostic Criteria for Research (World Health Organization Division of Mental Health, 1990). Because the CIS-R and the SCAN have independently developed algorithms these could generate discordant diagnoses due to algorithm differences. In order to eliminate any differences between existing algorithms a check on this was made by writing a new algorithm. We chose non-psychotic depressive episode or disorder (F32 or F33) because we found a difference in the handling of severity of depression between the field trial and published Diagnostic Criteria for Research (World Health Organization, 1993). Criterion items were identified easily in both interviews using the existing maps for both algorithms.

The SCAN and CIS-R standard score outputs use somewhat different items, rendering comparison less meaningful. A new depressive symptom score was generated. It included the CIS-R depressive symptom questions and corresponding SCAN item ratings used in our new ICD-10 depression algorithm. They were eight somatic syndrome items and nine general depression items with a theoretical maximum score of 17. Any difference between these two scores must be due to differences at the item level.

Analysis

Weighting of the survey data was not required because all SIFT positive subjects were asked to participate in re-interviews. This contrasts with previous two phase design comparison studies, in which respondents were selected in different proportions (Anthony *et al.* 1985). We did not have re-interview data on SIFT negative respondents, an important point that is considered in the discussion.

CIS-R diagnostic coverage is limited to those non psychotic disorders (ICD-10 F32 to F42) listed in the report of the national survey (Meltzer *et al.* 1995*b*). SCAN interviewers covered a wider range of diagnoses not covered by the CIS-R but these could not be included in analyses. For each ICD-10 diagnostic class common to the CIS-R and the SCAN 2 × 2 tables were created. Detailed diagnoses were non-hierarchical; thus all diagnoses generated were included in analysis tables. The CIS-R does not distinguish between single and recurrent episodes of depressive disorder (F32 and F33) so that these categories were combined as 'any depressive disorder'. The CIS-R does not classify Dysthymia (F34), a chronic form of depression: the additional label 'any case of depression' was used to compare CIS-R 'depressive disorder' with SCAN 'depressive disorder or dysthymia'. A grouping termed 'mixed anxiety and depression', available in the CIS-R only, was not analysed. Hierarchical groupings of specific disorders were also determined. The different types of anxiety disorder that included phobia were combined into a category 'any phobic anxiety'. All anxiety disorders and obsessive–compulsive disorder made up 'all neurosis excluding depression'. The final 'catch all' category of 'any ICD-10 disorder' encompasses all of the more specific disorders.

Statistics reported include tabulated frequencies that independent investigators can analyse further. Kappa (Cohen, 1968) was chosen as the primary, planned measure of concordance, taking account of the relative merits of the available measures of concordance (Shrout *et al.* 1987; Spitznagel & Helzer, 1987; McLeod *et al.* 1990). The sensitivity and specificity of the CIS-R was calculated using SCAN as the standard. Percentage agreement and the proportion of subjects for whom a positive diagnosis was recorded on whom both instruments agreed (the Index of Agreement) were calculated, as was the total percentage agreement. The 95% confidence intervals were calculated for all of the proportional values and kappa. To assess whether concordance is affected as the delay between interviews increases, subset analyses were carried out in two groups, with respondents divided at the median interval between the two interviews. Sensitivity analyses were carried out using the depression symptom scores.

RESULTS

Yield of paired interviews and results of sampling procedures

At the 1882 addresses assigned and visited, 1170 adult occupants were identified as eligible: of these, 863 agreed to participate with an initial SIFT by completing the CIS-R. Ninety-six declined to be interviewed. The reasons for not interviewing the remaining 211 were: failure to make contact, 90; sampling error, 27; objection or refusal by another occupant of the house or failure to make contact with an occupant, 94. We employ the term 'Refusal', as used by ONS, to identify eligible subjects not interviewed for any reason. Initial 'Refusal' to undergo a CIS-R interview was marginally greater in this study (25%) than in the national survey (20·4%).

The number of SIFT positive respondents based on the ⩾ 8 threshold was 387 (45%) of the 863 CIS-R interviews, which was far higher than in the national survey (25%). Twenty-nine of the 123 whose second interview was the CIDI refused the third with the SCAN. Of those eligible 61% achieved a second diagnostic interview within the time interval and 44% completed a third interview. The final yield was 205 SCAN interviews with adults who had completed a CIS-R. Because of the need to complete assessments within the planned time scale, some SCAN interviews were not achieved. Non-cooperation by the respondent was not the only reason. A detailed chart along with other details of the sampling (Brugha *et al.* 1997*b*) is available on request.

Description of sample and comparison with OPCS National Survey and with 1991 Census

We compared selected sociodemographic characteristics of the comparison sample with that of the national survey sample and National Census data. Apart from the differences in refusal rates, the proportions under each heading in the two surveys were reassuringly similar. In the comparison sample, the proportion of women and of urban dwellers was higher. As expected in Leicester, there were also increased numbers of persons from ethnic minorities. In general, we concluded that the aim to match the Leicester sample to that found in the national survey was achieved satisfactorily. Details of these comparisons can be requested (Brugha *et al.* 1997*b*).

Instrument performance

Table 1 sets out the frequency of each disorder, according to the presence or absence of the ICD-10 neurotic and depressive disorders covered in both instruments. Overall percentage agreement ranged from 0·7 to 0·9, but the percentage agreement calculated when excluding those who score negative in both tests ranged from 0 to 0·3. Concordance findings according to kappa are in the poor to very poor range (Table 1). The upper bounds of the 95% confidence intervals do not exceed 0·5 and the lower bounds are at or close to −0·1.

Table 2 sets out the sensitivity and specificity findings. In general the specificity of the CIS-R was very good. The very poor sensitivity of the CIS-R for SCAN phobic disorders is striking: 12 of the 13 cases of phobia on the SCAN were specific phobias (see Table 1), and the 13th was a case of social phobia (also misclassified by the CIS-R). Sensitivity for depression was also poor and overall the sensitivity of the CIS-R was unsatisfactory. Generalized Anxiety Disorder accounted for most CIS-R anxiety diagnoses and was rare in SCAN interviews. Obsessive–compulsive disorder was also rare under SCAN (three of 205 interviews) but more common under the CIS-R (11 cases). There were no psychosis cases on SCAN during the month.

Of interview pairs 88% were completed within the planned 14-day interval and 99% within 21 days. Subset analyses were carried out in two groups of respondents divided at the median interval between the two interviews. Ninety-five pairs of interviews (46·3%) were completed within 7 days. There was no trend in the tables for concordance to be better for this group than for those interviewed after a longer interval although there was a suggestion that problems might arise with depressive disorder re-assessed over longer intervals (tables available on request).

When we used our own algorithm for classifying CIS-R and SCAN depressive episode or disorder the diagnostic allocations were remarkably similar to those for the supplied ICD-10

Table 1. *SCAN by CIS-R comparison: frequencies and concordance according to kappa* (N = 205)

Test (CIS-R) by SCAN		True	False	False	True		
		negative	positive	negative	positive		
ICD-10 diagnosis	ICD-10 code(s)	(−/−)	(−/+)	(+/−)	(+/+)	kappa	95% CI
Any ICD-10 non-psychotic diagnosis	Any F32, F33, F40, F41, F42	137	33	18	17	0·25	0·10 to 0·40
Any anxiety disorder	Any F40 or F41	156	25	13	11	0·26	0·09 to 0·43
Any case of depression	Any F32 or F33	181	11	9	4	0·23	0·00 to 0·46
Any phobic anxiety	F40.00, F40.01, F40.1, F40.2	189	2	13	1	0·10	−0·11 to 0·30
All neurosis	Any F40, F41 or F42	149	29	14	13	0·26	0·10 to 0·42
Any non-phobic anxiety disorder	F41.00, F41.01, F41.1	163	28	7	7	0·21	0·04 to 0·38
F32, F33							
Acute mild depressive episode or disorder without somatic features	F32.00, F33.00	198	2	5	0	−0·01	−0·03 to 0·00
Acute mild depressive episode or disorder with somatic features	F32.01, F33.01	201	4	0	0	*	— —
Any acute mild depressive episode or disorder	F32.00, F32.01, F33.00, F33.01	194	6	5	0	−0·03	−0·04 to 0·01
Any acute moderate depressive episode or disorder	F32.10, F32.11, F33.10, F31.11	193	9	3	0	−0·02	−0·04 to 0·00
Any acute severe depressive episode or disorder	F32.2, F33.2	205	0	0	0	*	— —
Any acute or chronic case of depression	Any F32, F33 or dysthymia	184	13	6	2	0·13	−0·08 to 0·34
F40, F41, F42							
Any agoraphobic anxiety disorder	F40.00, F40.01	203	2	0	0	*	— —
Agoraphobia with panic disorder	F40.01	204	1	0	0	*	— —
Social phobias	F40.1	201	1	3	0	−0·01	−0·02 to 0·00
Specific (isolated) phobias	F40.2	193	0	12	0	*	— —
Any panic disorder	F41.00, F41.01	192	0	12	1	0·14	−0·10 to 0·37
Panic disorder (moderate)	F41.00	202	1	2	0	−0·01	−0·02 to 0·00
Generalized anxiety disorder	F41.1	170	34	1	0	−0·01	−0·03 to 0·01
Any obsessive–compulsive disorder	F42, F42.1, F42.2	192	10	2	1	0·12	−0·12 to 0·37

* Kappa cannot be calculated due to zero cells.

algorithms. Concordance, based on kappa, ranged from 0·16 (any case of depressive episode or disorder, F32 or F33) to 0·24, a non-statistically significant difference. Thus, allowance for identifable differences between the two published algorithms did suggest a difference, but one that was not sufficient to explain the poor concordance for depression.

When the depression symptom scores, based on the same criterion symptoms used in ICD-10, were compared the CIS-R score ranged from 0 to 15, with a mean score of 4·6 (95% CI 4·2 to 5·0). The SCAN depression score ranged from 0 to 10, with a mean score of 1·8 (95% CI 1·5 to 2·1). Significant bias was shown. The non-parametric correlation of the CIS-R and SCAN depression scores was 0·42 (Kendall's tau b; $P < 0{\cdot}01$; $N = 205$). We examined the effect of the CIS-R impaired functioning question by coding each depression symptom as zero if the respondent had no 'overall impairment'. The bias was completely eliminated, both scores having identical means and ranges and the correlation was almost the same (Kendall's tau = 0·39).

In order to examine whether agreement is affected by the use of categorical instead of dimensional data, cut points were applied to the depression scores. In order to correspond to the proportions in Table 1 for 'any depressive disorder' cuts were made at the 4th percentile on the SCAN score and at the 7th percentile on the CIS-R score. Thus, respondents above these cut-points were cases and those below non-cases. In order to compare 'like with like' statistically we used the Kendall's tau statistic in the resulting 2 × 2 table also. Agreement according to Kendall's tau was 0·38 (kappa was 0·35). When respondents without impaired functioning on the CIS-R were coded as non-cases agreement was similar (kappa = 0·24). We also examined cut-points at the 12th and 40th percentiles (on SCAN) with corresponding cut points on

Table 2. *SCAN/CSIR comparison: sensitivity and specificity of CIS-R disorders according to SCAN as the standard* (N = 205)

		Sensitivity		Specificity	
Diagnosis	ICD-10 code(s)	Value	95% CI	Value	95% CI
Any ICD-10 diagnosis	Any F32, F33, F40, F41, F42	0·49	0·31 to 0·66	0·81	0·75 to 0·86
Any anxiety disorder	Any F40 or F41	0·46	0·26 to 0·67	0·86	0·81 to 0·91
Any case of depression	Any F32 or F33	0·31	0·09 to 0·61	0·94	0·90 to 0·97
Any phobic anxiety	F40.00, F40.01, F40.1, F40.2	0·07	0·00 to 0·34	0·99	0·96 to 1·00
All neurosis	Any F40, F41 or F42	0·48	0·29 to 0·68	0·84	0·78 to 0·89
Any non-phobic anxiety disorder	F41.00, F41.01, F41.1	0·50	0·23 to 0·77	0·85	0·80 to 0·90
F32, F33					
Acute mild depressive disorder without somatic features	F32.00, F33.00	0·00	0·00 to 0·52	0·99	0·96 to 1·00
Acute mild depressive disorder with somatic features	F32.01, F33.01	*	— —	0·98	0·95 to 0·99
Any acute depressive disorder	F32.00, F32.01, F33.00, F33.01	0·00	0·00 to 0·52	0·97	0·94 to 0·99
Acute depressive disorder	F32.10, F32.11, F33.10, F33.11	0·00	0·00 to 0·71	0·96	0·92 to 0·98
Acute depressive disorder	F32.2, (no F33.2)	*	— —	1·00	0·98 to 1·00
Any depressive disorder	Any F32, F33 or dysthymia	0·25	0·03 to 0·65	0·93	0·89 to 0·96
F40, F41, F42					
Any agoraphobic anxiety disorder	F40.00, F40.01	*	— —	0·99	0·97 to 1·00
Agoraphobia with panic disorder	F40.01	*	— —	1·00	0·97 to 1·00
Social phobias	F40.1	0·00	— —	1·00	0·97 to 1·00
Specific/isolated phobias	F40.2	0·00	0·00 to 0·26	1·00	0·98 to 1·00
Any panic disorder	F41.00, F41.01	0·08	0·00 to 0·36	1·00	0·98 to 1·00
Panic disorder (moderate)	F41.00	0·00	0·00 to 0·84	1·00	0·97 to 1·00
Generalized anxiety disorder	F41.1	0·00	0·00 to 0·97	0·83	0·78 to 0·88
Any obsessive–compulsive disorder	F42, F42.1, F42.2	0·33	0·01 to 0·91	0·95	0·91 to 0·98

* Cannot be calculated due to zero cells.

the CIS-R score and obtained essentially similar agreement coefficients. Thus, although concordance remained poor there was nothing to suggest that the effect of using the full scores provided an improvement over the use of various cut-points applied within the same scores. Although these non-significant trends appear interesting, taken together, these sensitivity analyses do not explain sufficiently the poor concordance between the CIS-R and SCAN for depression.

We also compared agreements for individual SCAN and CIS-R depression symptoms and items. Overall agreement was very poor but some items performed better than others. Several individual items achieved kappa values better than 0·2: weight loss due to loss of appetite due to depression (kappa = 0·3); marked loss of appetite due to depression (0·4); early morning wakening (0·3); disturbed sleep (0·4). A diagnosis of depression requires the presence of two of three 'B' criteria in ICD-10: kappa for depressed mood was 0·3. For decreased energy or fatigueability kappa was 0·04; this symptom was coded present in 79% of the CIS-R interviews, but was rated as pathological in only 8% of SCAN interviews. For 'loss of interest or pleasure in activities that are normally pleasurable' kappa was 0·23. This criterion was endorsed on 6% of SCAN interviews and on 10% of CIS-R interviews. Other examples suggestive of strong bias were found: psychomotor retardation or agitation (47% of CIS-R respondents and only 1/205 positive SCAN ratings); ideas of self reproach or guilt (61 CIS-R respondents and 7 positive ratings on SCAN); lack of emotional reaction to events (19 CIS-R respondents and two SCAN ratings respectively).

Length and comparability of interviews

The mean number of minutes taken to conduct the CIS-R and the SCAN was 30 min and the median was 36 and 34 min respectively.

According to detailed qualitative comments (Brugha *et al.* 1997*b*) both the SCAN and CIS-R interviewers found that their interviews were very acceptable to respondents. CIS-R interviewers did wonder sometimes whether respondents understood what was being enquired about in the interview in the sections on anxiety, phobias, compulsions and in the final section on the overall effect of symptoms on functioning.

The depression sections seemed to be better worded and more clearly understood.

DISCUSSION

This purpose-designed study shows poor agreement between the SCAN and the CIS-R on the identification of a range of specific ICD-10 neurotic disorders (Table 1). A range of sensitivity analyses failed to cast doubt on this finding. We have no reason to think that the measurement difficulties found in the present study raise concerns for the use, in social surveys, of structured interviews for other aspects of social behaviour. Our concerns relate to clinical diagnostic assessments obtained in mental health surveys and how to make optimal use of existing survey data. Before discussing the possible reasons for and implications of this lack of concordance possible limitations to the study need to be considered.

In addition to attrition at the sampling stage, as expected, there was further attrition at the second and third interview stages. We attributed this to two factors: first, the time interval allowed between interviews was exceeded; and secondly, there were two follow-up interviews, making considerable demands on respondents.

Ideally, the order of administration of SCAN and CIS-R should have been randomized. Order effects could have been estimated. However, the design of the study required that all survey respondents underwent a CIS-R interview initially. In our comparison of the SCAN and CIDI, in which random ordering proved feasible, order effects were found. There was a trend for concordance to be poorer when, as in the present comparison, the clinical interview was preceded by the lay structured interview (Brugha *et al.* 1999*b*).

Agreement between the SCAN and the CIS-R must be assessed in relation to the test–retest reliability of each instrument. If the instrument cannot agree with itself it is unlikely to agree with a different instrument. Neither of these instruments has been subject to test–retest assessment under identical conditions, i.e. in the general population, testing the reproducibility of diagnostic classifications. SCAN interviews have been repeated with individual patients approximately a week apart, yielding excellent concordance findings for schizophrenia and for depressive episode (Wing *et al.* 1998). The reproducibility of diagnoses might be worse in the general population.

The effect of our two-phase design on concordance estimates may be difficult to quantify because CIS-R SIFT negatives were not assessed at all by the SCAN. If we assume that that all SIFT negative respondents are non-SCAN cases the kappa coefficients would be marginally better, but still very low. Similarly, if a small number of CIS-R SIFT negatives yielded positive SCAN diagnoses, the concordance coefficients would not have been greatly affected. Arguably, any such effect would have been reduced concordance. As a further precaution, sensitivity analyses were carried out simulating 'worst possible scenarios', in which we estimated the effects on the comparison study findings of incorrectly allocating as SIFT negative small numbers of 'cases'. This did not noticeably alter the findings either.

The important assumption that few if any SIFT negatives using the ⩾ 8 cut-point, included SCAN cases can be considered further. The SIFT positive percentage in the comparison study, using the ⩾ 8 cut-point, was found to be was appreciably greater than predicted by data gathered in the national survey: 45% of respondents were invited for a re-interview with SCAN. A two phase general population survey in Leicestershire of women, 3–6 months after childbirth, employed the GHQ-30 (screening for the upper quartile of GHQ scores) a far less sensitive screen (Brugha *et al.* 1998). It successfully picked up all but one of 25 SCAN cases of depression (Brugha *et al.* 1998). CIS-R SIFT negative respondents can include a small number of positive CIS-R-ICD-10 diagnoses (well under 1% in the national data): five did so (also well under 1%). None of these had depression. The diagnoses were: generalized anxiety disorder, social phobia and obsessive–compulsive disorder, but the concordance for these was close to zero. The final conclusions would not be noticeably different even if these same respondents had yielded perfect concordance with SCAN.

Our conclusion, therefore, is that although the concordance estimates for disorders may be biased, the amount of bias is almost certainly small and cannot alter significantly the clear conclusions of the study.

Explanations for poor concordance

The pattern of distribution of disorders in the general population might also explain much of our findings. In contrast to clinical populations, disorders close to their defined thresholds are likely to outnumber those that are well above it, This is clearly illustrated in the distribution of total CIS-R scores reported in the national survey (see Fig. 4.1 in Meltzer *et al.* 1995*a*). The identification of positives is likely to be more difficult where most respondents are negative and most positives are borderline. There is an impression from the data presented here that the less frequent conditions have poorer levels of concordance. Slightly better levels of concordance were found when specific disorders were grouped together into overall ICD-10 categories (Table 1). We had expected to find, therefore, that the threshold for depression is critically important in a community sample. However, in the case of depression at least, this expectation was not born out. Lack of concordance may also be affected by coding and rating errors at the criterion (i.e. symptom) level, by unintended differences in algorithms and by the effect on concordance of using binary, ordinal or interval level data. Further sensitivity analyses were carried out also in order to examine the effects of these.

Differences in criterion or symptom ratings

Marked discrepancies between the two interviews were found at the individual symptom level despite the fact that many are not particularly uncommon. For example, of 162 respondents who replied 'yes' to the CIS-R question on fatigue only 16 were rated present by the SCAN interviewer. Patients in contact with psychiatric services may 'learn' the accepted meaning of clinical terminology, but this degree of shared comprehension cannot be assumed in general population responders. This would affect the CIS-R much more than the SCAN in which elucidatory questioning is mandatory. In developing the CIS-R as a structured, lay diagnostic measure, the flexible and clinical approach of the earlier CIS (Goldberg *et al.* 1970) has been lost.

Symptom scores

Analyses of summed symptom scores may help to identify the extent that poor concordance is due to the effect of large numbers of cases close to the threshold for disorder, because this does not affect symptom scores. Our sensitivity analyses with depression do not support this argument. Much more striking, though, CIS-R depression scores were significantly higher than those derived from the SCAN. This implies that ratings based on the CIS-R will impel individuals towards the thresholds for recognizing disorders. This corresponds to the observation that the overall agreement for the individual criteria is as bad as that for the diagnosis of depression. However, the score bias may have been due to the key factor of impairment of functioning, although by eliminating this through the inclusion of a single question the score correlation may suffer.

Algorithm differences

The existence of algorithm errors is receiving recognition from other researchers (Marcus & Robins, 1998); our work shows that this too may make a small contribution to discrepancies when comparisons are made between diagnostic interviews. But applying a single algorithm to the depression symptom data showed that this did not explain sufficiently the poor concordance for 'depressive episode or disorder'.

Other sources of error

Unlike a cut-point on a score, diagnostic rules may depend unduly on the validity of individual criteria. When compared to the presence, or absence, of a diagnosis or of an individual depression item the concordance between SCAN and the CIS-R remained poor but possibly slightly better when the full range of depressive symptoms was used. Probabilistic models that set different thresholds on criteria (Surtees *et al.* 1997) may be another way of addressing this.

Implications for the use of epidemiological data

Our findings point to a key message: findings from large-scale surveys, using the CIS-R, should take account of direct comparisons with clinically based assessments if they are to be meaningfully interpreted. With this proviso, the data obtained in the first national survey of psychiatric morbidity (Jenkins *et al.* 1997*b*) is of inestimable value. For example, our findings suggest that the prevalence estimates of depressive disorder in the first adult national survey (Jenkins *et al.* 1997*b*), to an extent, are over-estimates. An important and treatable group of specific anxiety disorders, panic and phobic disorder are frequently misclassified, although many of the phobias may be specific phobias of less clinical significance. The two- to three-fold difference in prevalence rates, for example for phobia, panic disorder and obsessive–compulsive disorder, would pose problems for national planning and allocation of resources for specific disorders if the additional information provided by the comparison with SCAN were omitted. But the comparison data may allow some translation of prevalences found with the CIS-R. Details will be published elsewhere of crosswalk analyses necessary to adjust estimates obtained from the national survey (Brugha *et al.* 1997*a*). Such analyses may help to establish quantitative estimates of the information limits of data gathered in national surveys using the CIS-R, which may be of practical use and importance to service commissioners and planners.

While the present study suggests that estimates of the prevalence of mental disorders, derived from large scale surveys, are not very wide of the mark, problems concerning the accuracy of case identification must also be considered. This may reflect on two areas of importance to epidemiologists: population needs assessment and the estimation of risk.

Assessment of need

It is clear that there are differences between lay and semi-structured clinician assessments in identifying which individuals are most clearly cases (Table 1) and therefore likely to be in need of health care interventions. In a local community survey of mental health needs (Bebbington *et al.* 1997), using detailed clinician reappraisals of need, it was found that diagnosis based on SCAN-ICD-10 is only an approximate indication of needs for treatment. But, if diagnosis itself is crude and inaccurate, it becomes even more suspect as an indicator of need. The comparison study suggests that an important and treatable group of disorders, panic disorder and phobia, are greatly underestimated by the CIS-R, while cases of obsessive–compulsive disorder are being incorrectly identified. In further work we hope that by incorporating SCAN data it will be possible to reduce the misclassification of cases. In the meantime results relating to service use and the very low treatment take up rates reported from the national survey in relation to specific neurotic disorders in the existing survey (Meltzer, *et al.* 1995*c*; Bebbington *et al.* 1998) will need to be interpreted with caution. Similarly, the design of future surveys will require greater use of clinically-based assessments, possibly with the SCAN, and more detailed studies of health care needs.

Estimation of risk

Studies of the association between potential risk factors and disorder rely either on accurate identification of cases or on unbiased estimates based on dimensional scales. If inaccuracy were predominantly the result of cases clustering around recognition thresholds, the effect on risk identification would be small, as cases and non-cases close to the threshold are likely to resemble one another. However, the finding that this is not the main factor is worrying.

Improving structured diagnostic measures

Information on sources of bias could be used to inform the design of structured questionnaires. By examining coding errors at criterion level it may be possible to develop new wordings of questions, thus leading to greater accuracy. Improved question wording might also result from cross-checking by trained clinicians of the respondents' own understanding of the meaning of structured questions, similar to the use of the technique known as cognitive interviewing (Jobe & Mingay, 1990). Both approaches might also help to identify criteria that it may be impractical to deal with accurately and efficiently with the use of structured questions. For these a better solution might be to collect verbatim descriptions (vignettes) for later rating by a clinically experienced assessor using, for example, SCAN glossary definitions. Kessler has provided a recent, more extensive, discussion of some of the available options (Kessler, 1999).

Future use of SCAN and other clinical measures

SCAN was used to re-interview a random sample of all respondents, who had completed a CIS-R interview in the recent national survey of psychiatric morbidity among prisoners in England and Wales (Singleton *et al.* 1998). This is a good example of how a clinical assessment can be incorporated into a large-scale survey in which a structured interview is the core measure of neurotic disorders. However, prevalence rates are lower in the general population than in the prisons and larger samples need to be assessed. To conduct a sufficient number of clinical assessments would require more suitably trained interviewers.

It is possible that lay, clinically inexperienced interviewers, could use semi-structured clinical interviewing methods, such as the PSE or SCAN. But they would have to be specially trained (Brugha & Nienhuis, 1998; Brugha *et al.* 1999*c*). The fore-runner of the SCAN (PSE-9) has been used by lay interviewers in three surveys of the general population, but only to assess non-psychotic disorders (Sturt *et al.* 1981; Dean *et al.* 1983; Rodgers & Mann, 1986). We have argued that SCAN used by trained clinicians is a more valid instrument than structured questionnaires (Brugha *et al.* 1999*a*). SCAN used by lay interviewers, however experienced, may be less valid. This is an important empirical question. Nevertheless, we now believe that lay interviewers of appropriate academic background and survey experience could be trained in the use of SCAN. They would need to undertake an extended course of guided experience in clinical examination of patients with specific psychotic and neurotic disorders. Their use of the SCAN must then be compared with that of experienced SCAN-trained clinicians. Since completing the present study we have evaluated the feasibility of training experienced survey interviewers to conduct SCAN interviews in psychotic and neurotic patients (Brugha *et al.* 1999*c*). Our initial findings in a clinical sample are highly encouraging.

Concluding comment

We still do not have a feasible survey measure of neurotic psychiatric disorder of acceptable validity that can be used alone in large-scale general population surveys. Greater accuracy is required to evaluate the appropriate use of health care (Leeman, 1998; Regier *et al.* 1998). Different approaches are warranted (Brugha *et al.* 1999*a*). Progress should be based either on substantially improved structured measures or on the incorporation of systematic, clinically-based measures, concurrently within large-scale national surveys, as in the recent prison survey (Singleton *et al.* 1998). Further efforts at 'fixing' existing measures may no longer be regarded as sufficient.

Given the considerable differences in the interviewing methods used, it is not surprising that these two approaches to symptom and diagnostic assessment generate different results in a general population sample. The differences in case identification and prevalence were conspicuous, and arise apparently from rating differences at the item level. We would argue from our results that data gathered in large-scale surveys must and can be interpreted in an informed and cautious manner. Existing survey findings with the CIS-R are useful provided they are interpreted in the light of the present findings.

This study was funded by a research grant from the Statistics Division of the Department of Health, London.

In addition to acknowledging the advice and assistance of the many national and international consultants listed below, we are grateful to the following for their help in conducting the survey: Ann Klepacz and Criag Norman of OPCS who developed and provided the first of the structured lay interviewing courses in Leicester; Anne Farmer for assisting with the CIDI training; Sheila Goward of the OPCS survey field-work team for her assistance with training our field workers and the initiation of the pilot and field-work stages; the sampling division at OPCS; James Gray of the Social Survey Division of OPCS who produced the computer assisted version of the CIS-R; Andy Stoker of the Division of Child Psychiatry Leicester for producing the CIS-R-ICD-10 diagnostic software from the algorithm supplied by OPCS; and, to the householders of Leicestershire who so generously gave of their time when we asked them to complete the study interviews.

The following participated as field work interviewers: 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.

Additional UK advisers consulted from the design stage were: G. Lewis, A. Farmer, D. R. Jones, D. D. Stretch and P. Surtees. International advisers who provided detailed advice at the design stage were: R. Kessler, A. Romanoski and W. Compton. Comments gratefully acknowledged since the initial communication of results were received from: R. Kessler, B. Üstün, N. Sartorius, G. Dunn, C. H. Brown, A. S. Henderson and A. Angold.

REFERENCES

Andrews, G., Peters, L., Guzman, A. M. & Bird, K. (1995). A comparison of two structured diagnostic interviews: CIDI and SCAN. *Australian and New Zealand Journal of Psychiatry* **29**, 124–132.

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.

Bebbington, P. E., Marsden, L. & Brewin, C. R. (1997). The need for psychiatric treatment in the general population: the Camberwell Needs for Care survey. *Psychological Medicine* **27**, 821–834.

Bebbington, P. E., Brugha, T., Jenkins, R., Lewis, G., Farrell, M. & Meltzer, H. (1998). Neurotic disorders and the use of services: a report from the National Survey of Psychiatric Morbidity (In preparation).

Brugha, T. S. & Nienhuis, F. J. (1998). *SCAN-SF. A Survey Form of the Present State Examination and SCAN: Supplementary PSE and SCAN Introductory Training Manual for Lay Interviewers*. World Health Organization SCAN Training Centre: Leicester.

Brugha, T. S., Teather, D., Wills, K. M., Kaul, A. & Dignon, A. (1996). Present State Examination by microcomputer: objectives and experience of preliminary steps. *International Journal of Methods in Psychiatric Research* **6**, 143–151.

Brugha, T. S., Jenkins, R., Bebbington, P., Meltzer, H. & Taub, N. A. (1997*a*). The scope for increasing the usefulness of population based epidemiological information on need. Conference abstracts: WPA Section of Epidemiology and Community Psychiatry, 19 October 1997, Sydney.

Brugha, T. S., Bebbington, P. E., Jenkins, R., Meltzer, H., Taub, N. A., Janas, M. & Vernon, J. (1997*b*). Cross Validation of the Lay Diagnostic Instruments Used in the Great Britain National Survey of Psychiatric Morbidity With Instruments Used in National Surveys Abroad. A comparison in a household survey of the Clinical Interview Schedule Revised, the Composite International Diagnostic Interview and the Schedule for Clinical Assessment in Neuropsychiatry as instruments used to measure the point prevalence of common psychiatric diagnoses in the general population. Supplementary and Final Report, 1995–1997, to the Department of Health, London. Department of Psychiatry, University of Leicester (Mimeo): Leicester.

Brugha, T. S., Sharp, H. M., Cooper, S. A., Weisender, C., Britto, D., Shinkwin, R., Sherrif, T. & Kirwan, P. H. (1998). The Leicester 500 Project. Social support and the development of postnatal depressive symptoms, a prospective cohort survey. *Psychological Medicine* **28**, 63–79.

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., Jenkins, R., Taub, N. A., Meltzer, H. & Bebbington, P. (1999*b*). The clinical validity of an international psychiatric diagnostic interview used in general population prevalence surveys. (Submitted for publication.)

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**, 703–711.

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

Dean, C., Surtees, P. G. & Sashidharan, S. P. (1983). Comparison of research diagnostic systems in an Edinburgh community sample. *British Journal of Psychiatry* **142**, 247–256.

Der, G. & Wing, J. K. (1992). *The ICD 10 Program icd10*. MRC Social Psychiatry Unit: London.

Farmer, A. E., Jenkins, P. L., Katz, R. & Ryder, L. (1991). Comparison of CATEGO-derived ICD-8 and DSM-III classifications using the composite international diagnostic interview in severely ill subjects. *British Journal of Psychiatry* **158**, 177–182.

Goldberg, D. P., Cooper, B., Eastwood, M. R., Kedward, H. B. & Sheperd, M. (1970). A standardised psychiatric interview for use in community surveys. *British Journal of Preventive and Social Medicine* **24**, 18–23.

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.

Janca, A., Robins, L. N., Bucholz, K. K., Early, T. S. & Shayka, J. J. (1992). Comparison of Composite International Diagnostic Interview and clinical DSM-III-R criteria checklist diagnoses. *Acta Psychiatrica Scandinavica* **85**, 440–443.

Jenkins, R., Bebbington, P., Brugha, T., Farrell, M., Gill, B., Lewis, G., Meltzer, H. & Petticrew, M. (1997*b*). 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.

Jobe, J. B. & Mingay, D. J. (1990). Cognitive laboratory approaches to designing questionnaires for surveys of the elderly. *Public Health Reports* **105**, 518–524.

Kendler, K. S. & Gardner, C. O. (1998). Boundaries of major depression: an evaluation of DSM-IV criteria. *American Journal of Psychiatry* **155**, 172–177.

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.

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

Kovess, V., Sylla, O., Fournier, L. & Flavigny, V. (1992). Why discrepancies exist between structured diagnostic interviews and clinicians' diagnoses. *Social Psychiatry and Psychiatric Epidemiology* **27**, 185–191.

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.

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. Office of Population Censuses & Surveys Social Survey Division. (1995*a*). *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 Stationary Office: London.

Meltzer, H., Gill, B., Petticrew, M. & Hinds, K. Office of Population Censuses & Surveys Social Survey Division. (1995*b*). *OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 1: The Prevalence of Psychiatric Morbidity among Adults Living in Private Households. Algorithms for Production of ICD-10 Diagnoses of Neurosis from the CIS-R.* OPCS Surveys of Psychiatric Morbidity in Great Britain. Her Majesty's Stationary Office: London.

Meltzer, H., Gill, B., Petticrew, M. & Hinds, K. Office of Population Censuses & Surveys Social Survey Division. (1995*c*). *OPCS Surveys of Psychiatric Morbidity in Great Britain. Report 2:*

Physical Complaints, *Service Use and Treatment of Adults with Psychiatric Disorder*. OPCS Surveys of Psychiatric Morbidity in Great Britain. Her Majesty's Stationary Office: London.

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.

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. & Trowle, 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.

Rodgers, B. & Mann, S. A. (1986). The reliability and validity of PSE assessment by lay interviewers: a national population survey. *Psychological Medicine* **16**, 689–700.

Shrout, P. E., Spitzer, R. L. & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry* **44**, 177.

Singleton, N., Meltzer, H., Gatward, R., Coid, J. & Deasy, D. Office for National Statistics Social Survey Division. (1998). *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*. ONS Surveys of Psychiatric Morbidity in Great Britain. Her Majesty's Stationary Office: London.

Spengler, P. A. & Wittchen, H. U. (1988). Procedural validity of standardized symptom questions for the assessment of psychotic symptoms – a comparison of the DIS with two clinical methods procedural validity of standardized symptom questions for the assessment of psychotic symptoms – a comparison of the DIS with two clinical methods. *Comprehensive Psychiatry* **29**, 309–322.

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

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. (1987). Charlie Brown and statistics: an exchange (correspondence on use of 'Y statistic' for comparing inter rater reliability). *Archives of General Psychiatry* **44**, 194–195.

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.

Surtees, P. G., Wainwright, N. W., Gilks, W. R., Brugha, T. S., Meltzer, H. & Jenkins, R. (1997). Diagnostic boundaries, reasoning and depressive disorder. II. Application of a probabilistic model to the OPCS general population survey of psychiatric morbidity in Great Britain. *Psychological Medicine* **27**, 847–860.

Wilson, P. & Elliott, D. (1987). The evaluation of the Postcode Address File as a sampling frame and its use within OPCS. *Journal of Royal Statistical Society* **150**, 230–240.

Wing, J. K., Cooper, J. & Sartorius, N. (1974). *Measurement and Classification of Psychiatric Symptoms*. Cambridge University Press: Cambridge.

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). *Diagnosis and Clinical Measurement in Psychiatry*. *A Reference Manual for SCAN/PSE-10*. 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., Robins, L. N., Cottler, L. B., Sartorius, N., Burke, J. D. & Regier, D. (1991). Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI). The Multicentre WHO/ADAMHA Field Trials. *British Journal of Psychiatry* **159**, 645–658.

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

World Health Organization Division of Mental Health (1990). *ICD-10. Chapter V Mental and Behavioural Disorders* (*including disorders of psychological development*)*: Diagnostic Criteria for Research* (*May 1990 Draft for Field Trials*). World Health Organization: Geneva.

World Health Organization Division of Mental Health (1992). *WHO SCAN Advisory Committee. SCAN Schedules for Clinical Assessment in Neuropsychiatry*, *Version 1.0*. *Schedules for Clinical Assessment in Neuropsychiatry*. *Distribution from Training Centres*. World Health Organization: Geneva.