Psychological Medicine, 1999, **29**, 1013–1020. Printed in the United Kingdom

EDITORIAL

A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population[1]

THE QUEST FOR PREVALENCE

Psychiatric case-identification in general populations allows us to study both individuals with functional psychiatric disorders and the populations from which they come. The individual level of analysis permits disorders to be related to factors of potential aetiological significance and the study of attributes of the disorders that need to be assessed in non-referred populations (an initially scientific endeavour). At the population level valid case identification can be used to evaluate needs for treatment and the utilization of service resources (a public health project). Thus, prevalence is of interest both to scientists and to those responsible for commissioning and planning services (Brugha *et al*. 1997; Regier *et al*. 1998). The quality of case identification techniques and of estimates of prevalence is thus of general concern (Bartlett & Coles, 1998).

Structured diagnostic interviews were introduced into general population surveys in the 1970s as a method 'to enable interviewers to obtain psychiatric diagnoses comparable to those *a psychiatrist would obtain*' (Robins *et al*. 1981). The need to develop reliable standardized measures was partly driven by an earlier generation of prevalence surveys showing rates ranging widely from 10·9% (Pasamanick *et al*. 1956) to 55% (Leighton *et al*. 1963) in urban and rural North American communities respectively. If the success of large scale psychiatric epidemiological enquiries using structured diagnostic interviews and standardized classifications is measured in terms of citation rates it would seem difficult to question. But the development of standardized interviews of functional psychiatric disorders has not solved this problem of variability: the current generation of large scale surveys, using structured diagnostic interviews and serving strictly defined classification rules, have generated, for example, 12-month prevalence rates of major depression in the US of 4.2% (Robins & Regier, 1991) and 10·1% (Kessler *et al*. 1994). This calls into question the validity of the assessments, such that we must reopen the question of what they should be measuring and how they should do it.

WHAT IS PSYCHOPATHOLOGY?

The nineteenth and early twentieth centuries saw considerable progress in the description of mental experience, with particular impetus coming from German-speaking psychiatrists. Mental experiences were categorized as symptoms, by analogy with the way physicians handle their patients' physical experiences, and as a way of distinguishing between putative mental diseases. Our current classifications in psychiatry all carry specific reference to symptoms defined in this way, and this legacy still forms the basis of diagnosis.

Hand-in-hand with the development of psychopathological descriptions, the principles of clinical assessment now known as 'the mental state examination' were established. This is a demanding list of requirements: a sound grasp of clinical psychopathology, together with extensive experience gained from interviewing patients with a wide variety of neurotic and psychotic symptoms. In order

[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.

to be able to recognize them in patients, the student must learn to identify symptoms (such as delusion, hallucination, obsessional rumination, irrational fear of harmless stimuli and excessive feelings of guilt) in a standard manner. Techniques of cross-questioning are learnt from observing experienced practitioners and trying them out in practice. It is those techniques of cross-examination that practitioners regard as central to the evaluation of psychopathological features. A clear historical link can be traced to the modern concepts of psychiatric disorders incorporated in the standard classifications. Thus, if there is a gold (or reference) standard of psychopathology assessment, founded on this historical legacy, it embodies these elements: knowledge and experience of abnormal states of mind, and the skill to elicit them and to make judgements as to their presence and significance. Standardized interviews of the semi-structured variety endeavour to be faithful to the legacy in its entirety but in a systematic way; fully structured measures by definition must omit both the element of clinical judgement and the element of flexible cross questioning. But are these two elements essential and is there a price to be paid for simplification by omission?

THE DRIVE FOR STANDARDIZATION

The problem for researchers with the examination of the mental state is that it has the qualities of an art, not a technique. It is extremely difficult to standardize such a process, and yet this is a prerequisite to the comparability that makes research possible. There have been two general approaches to eliminating the idiosyncrasies of individual clinicians.

The first technique is the semi-structured standardized interview, an attempt to encapsulate the essence of clinical cross-examination, while limiting its potential vagaries. The interviewers for these clinical measures are usually psychiatrists or clinical psychologists. They undergo specific training in the use of the instrument. They are provided with standard wording of the questions to be asked, but, where the answers are unclear, are encouraged to follow-up with a freeform enquiry. This requires additional probe questions using words and phrases that respondents can understand, often incorporating the respondent's preferred verbal descriptions of their complaints. The interview is flexible. It requires of the respondents no comprehension or knowledge of psychopathological terminology. However, the examiner must query and clarify any use of 'jargon' or idiosyncratic usage by the respondent. Each symptom is defined. In the case of the archetypal example of such an instrument, the Present State Examination (PSE; Wing *et al.* 1974) now incorporated into the Schedules for Clinical Assessment in Neuropsychiatry (SCAN; Wing *et al.* 1990; World Health Organization Division of Mental Health, 1992), definitions are embodied in an explicit glossary of symptoms and abnormalities of behaviour. Examiners compare respondents' descriptions of their mental experiences with the glossary (criterion) definitions. In other words, they match the patient's experience with the conception that lies behind individual symptoms. Once this information has been gathered, immediate human judgement is no longer required to allocate the respondent to a diagnostic category: ratings of symptoms are submitted to a computer algorithm that serves the major standardized classifications. Thus, the interviewers do not make a diagnosis; they merely decide on the ratings of individual symptoms.

The alternative to these semi-standardized interviews is one in which the standardization is effectively total (Spengler & Wittchen, 1988). The best known of these instruments are the Diagnostic Interview Schedule (DIS; Robins *et al.* 1981) and the Composite International Diagnostic Interview (CIDI; Robins *et al.* 1988). A more recent development is the revised Clinical Interview Schedule (CIS-R; Lewis *et al.* 1992) that grew out of an earlier semi-structured interview (Goldberg *et al.* 1970). Survey interviewers using these completely structured measures are not expected to have any clinical background or experience. Indeed, this is the whole point, as such interviewers are less expensive than trained clinicians. They undergo general training in survey interviewing, and training and familiarization in the specific measure. The interviewer recites as written pre-worded questions that ask about symptoms (Spengler & Wittchen, 1988) and further questions deal with the occurrence, frequency, duration and severity of the symptoms in the assessment period covered. If the respondent replies 'yes' to a given question, the interviewer must

code this endorsement. The interviewer is not permitted to make judgements, apart from one: has the respondent understood the question just asked? If not, the question is repeated verbatim. Much of the sophistication of these questionnaires lies in the use of preambles, the careful phrasing of questions in such a way as to attempt to explain the meaning of these terms to the respondent, and efficient follow-up in response to particular answers. Both the CIDI and CIS-R are available in fully automated versions (Lewis *et al*. 1988; World Health Organization, 1993*a*) exemplifying their fully structured modus operandi, and can be self-completed or administered by means of a 'Computer Assisted Personal Interview' (CAPI).

In order to use a fully structured interview, experienced health survey interviewers need only a few hours of familiarization with the rationale of the interview and the meaning of questions, thus further reducing survey costs. However, the extent to which all such interviews are fully structured can be difficult to infer clearly from the literature: DIS training courses can take at least one week (Robins *et al*. 1981) and authorities seem to differ on the extent to which interviewers are required (or permitted) to exercise discretion in the use of additional probe questions and indeed ratings (Robins *et al*. 1981; Spengler & Wittchen, 1988), which suggests that some experts do acknowledge reservations about the adequacy of totally structured interviews. Semi-structured interviewers require training commensurate with the greater skills required by the method, ranging from a week for the experienced clinician (Wing *et al*. 1990) to a month for the professional survey interviewer without a clinical background (Brugha *et al*. 1999*a*).

THE DIFFERENCE BETWEEN THE TWO METHODS OF SYSTEMATIC INTERVIEWING

Semi-structured clinical assessment and lay-administered structured questionnaires both approach the problem of cost-effectiveness. In the semi-structured assessment, the clinical experience of the interviewer 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. It is conceivable to circumvent the interviewer's clinical judgement in this process, by considering all possible responses and framing follow-up questions to deal with them. However, this could become an exhaustive and exhausting procedure (Brugha *et al*. 1996). The cost of relying on clinical judgement may be a loss of control over standardization and therefore a potential reduction in reliability (Lewis *et al*. 1992). The reliability of these interviews (also called reproducibility) is maximized by standardizing coverage, providing a structure (with rules about cut-off procedures) and training interviewers to criteria that are embodied in common standards defined in a glossary. The fully structured lay questionnaire seeks cost-effectiveness by eliminating the need for clinically experienced interviewers. However, this is done at the possible expense of validity (also termed precision), since there is no clinical evaluation of responses before a symptom is coded present or absent. Moreover, the questionnaire must be restricted to an acceptable and feasible length, and this too will restrict validity. These trade-offs between feasibility, cost-effectiveness, reliability and validity are inevitable.

The approach taken in semi-structured instruments such as the Present State Examination (Wing *et al*. 1974) approximates most closely to the full range of key elements of clinical assessment discussed above, but in a standardized manner that is feasible in population surveys. A few investigators have argued that, at this stage of development, it is not appropriate to choose between the two approaches to interviewing. They have therefore argued against reporting sensitivity and specificity, as this requires that one measure be the standard against which to judge the other (Andrews *et al*. 1995). However, most authorities probably agree that a systematic clinical assessment is the standard by which to assess lay measures (Spitzer, 1983; Anthony *et al*. 1985; Helzer *et al*. 1985; Romanoski *et al*. 1988; Wittchen, 1994; Kessler *et al*. 1998). In clinical populations, both methods have been shown to reproduce findings reliably when re-administered by an independent interviewer after an interval (Wing *et al*. 1998).

It is important to remember the ways in which these instruments resemble each other. Both follow standard diagnostic classification rules, and these are applied by means of computer algorithms making use of ratings of the specific items required by the classification. Furthermore, a high level of consensus has been achieved in the development and acceptance of classification rules (World Health Organization, 1993*b*; American Psychiatric Association, 1994). Comparisons between different methods of assessment using different classification systems would not be of value as the multiple sources of variation would make it impossible to draw unambiguous inferences from the results. In the next section we discuss comparisons between different interviewing methods in which the same classification system is used – that is in which the same items are covered in both interviews. Therefore, differences between the interviews being compared must be due to measurement (i.e. coding or rating) and not classification differences.

COMPARISONS OF STRUCTURED AND SEMI-STRUCTURED INTERVIEWS THAT USE THE SAME CLASSIFICATION SYSTEM

Investigators have recognized the need for validity studies in the general population, the setting for which structured interviews are primarily intended and in which they should be compared with clinical instruments (Wittchen, 1994). Compared with the clinical populations in which such instruments have generally been shown to be reliable (Page, 1991), the low prevalence of disorders in the general population presents a greater challenge (Spitzer, 1983). A few comparisons of structured and clinical interviews have been made in samples of householders sufficiently large to assess validity and drawn randomly from a population and not selected for specific diagnoses. Unfortunately, none of the instruments have been so compared against themselves (for test–retest reliability) in such settings except at the individual item level (Cooper *et al.* 1977), an essential estimate of the optimum level attainable of reproducibility.

Anthony and his colleagues (Anthony *et al.* 1985) compared the Diagnostic Interview Schedule (DIS) with a clinical measure conducted by trained psychiatrists and based on the PSE (Romanoski *et al.* 1988). This comparison involved the assessment of 802 subjects in a household sample in Baltimore. The DIS performed very poorly as a measure of neurotic and psychotic psychiatric disorders and hence of their underlying symptoms in the month prior to interview (Anthony *et al.* 1985). In particular, the lay and clinical measures largely failed to agree on which individuals were cases of particular disorders.

In a second study based on a Chicago household survey, the DIS was compared with the SADS-L (Spitzer *et al.* 1978). This is a semi-structured diagnostic interview administered by clinicians, and in this case by clinically experienced psychiatric social workers (McLeod *et al.* 1990). Substantial discrepancies were found in the Research Diagnostic Criteria diagnosis of depression over a period of 6 months. These were attributed to recall error and particularly to inconsistent reports of episode timing within the 6 month period covered by the two instruments.

The Composite International Diagnostic Interview (CIDI; Robins *et al.* 1988) has been the subject of an enormous amount of development work and has been widely used. Its reliability and validity has been studied and discussed by a number of investigators (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 clinical populations, good to excellent reliability has been reported for most diagnostic sections (Wittchen, 1994). Unlike the DIS, however, almost all projects make use of non-standard versions of the CIDI (Wittchen, 1996; Wittchen *et al.* 1998), a factor that may contribute to some of the difficulties encountered in trying to achieve standardized prevalence estimates.

Only the University of Michigan version of the CIDI (UM-CIDI) (Wittchen *et al.* 1995, 1996) has been compared with a clinical diagnostic interview in the general population. This was used in the US National Comorbidity Survey (Kessler *et al.* 1994) for which the high reported estimates of the prevalence of disorder have been the subject of significant concern (Regier *et al.* 1998). The UM-CIDI has only been compared with individual diagnostic modules of a clinician-administered

structured diagnostic interview, the Structured Clinical Interview for DSM-III-R (SCID; Spitzer *et al*. 1992). This reappraisal study was conducted in samples of householders selected for each particular diagnosis by means of the UM-CIDI stem (screening) questions (Wittchen *et al*. 1996). Therefore, procedural validity (concordance between the UM-CIDI and SCID) for social phobia and agoraphobia (Wittchen *et al*. 1996) and for other neurotic disorders could not be accurately estimated because of cost constraints on the study scope and design (Kessler *et al*. 1998). For this reason, it cannot be equated with the comparison studies quoted earlier. In particular, the design used will tend to underestimate the proportion of missed true cases: indeed the authors were cautious in pointing out that only upper bound estimates could be produced for concordance, sensitivity, specificity and negative predictive value; therefore, it is not surprising that these estimates may appear to be more satisfactory than in the earlier evaluations of structured interviews.

The Revised Clinical Interview Schedule (CIS-R; Lewis *et al*. 1992), a fully structured interview, was designed to detect ICD-10 neurotic disorders. Unlike the DIS and CIDI, which set out to measure disorders throughout the life course, the CIS-R covers symptoms in the past week, and only extends to longer periods of up to 4 weeks prior to interview where these are required for certain classification rules. Unlike the CIDI there is only one version of the CIS-R, thus ensuring comparability between studies. A general population survey was conducted to compare the SCAN and CIS-R in 205 householders at moderate risk of psychiatric disorder: only 2 of 8 SCAN cases of depressive disorder were correctly classified by the structured interview; moreover, 13 individuals were also miss-classified as cases by the structured interview (Brugha *et al*. 1997, 1999*b*). These discrepancies are consistent with the other studies quoted earlier.

All-in-all, the general population comparison studies quoted here concur in suggesting that a high proportion of important and potentially treatable disorders, such as depression, panic disorder and phobia, will be incorrectly identified by structured interviews.

IS THERE A WAY FORWARD?

When standardized instruments were introduced, it was on the assumption that they would deal with the twin problems of inconsistent case identification and variable prevalence. This assumption has proved over-optimistic. The first of these may be an attribute of 'imperfect' instruments, but the problem may be compounded by the convention of using classification systems developed in secondary care that emphasize rare symptoms and disorders, a problem that might be mitigated by the use of primary care versions (Pincus & McQueen, 1997). If the problem is that of the use of 'imperfect' instruments, this could be because the essential elements of flexible cross questioning and rating judgement are critically important in interviews with community respondents who may not share the interview designer's understanding of clinical terms. The general population comparison study findings are suggestive of this explanation.

The effect of misrecognition on, for example, the sociodemographic attributes of disorder will not usually be significant. Of greater concern is the particular importance to public health epidemiologists of accuracy of case identification for population needs assessment. It is clear from the published studies that there are substantial differences between lay survey interviews and semi-structured clinician assessments in terms of identifying which individuals are most clearly cases and therefore likely to be in need of health care interventions. This in turn must bias population level estimates of the resources that should be committed to addressing these needs.

The problem of variable prevalence is potentially much more damaging to the purposes of epidemiology – it makes comparisons between different studies much more difficult to interpret. This is particularly important at present because psychiatry has reached a new stage of globalization. Thus, epidemiologically-based prevalence data are used to derive estimates of the total amount of disability occurring in different populations. The Global Burden of Disease Report has drawn attention to the importance of psychiatric diseases as a 'highly significant component of global disease burden' (Murray & Lopez, 1996). Much of the data available for national and

international comparisons are likely to be based on the widespread use of structured diagnostic interviews in epidemiological surveys in countries both within and outside America and Europe (Weissman *et al.* 1996).

The design of future surveys will need to take the difficulties highlighted here into consideration. In our view this will require substantially greater use of standardized clinically-based assessments that employ the key elements of psychopathology assessment referred to earlier. At the same time, reference measures must also be shown to be reliable and reproducible following independent assessments. We believe that the considerable achievements of our current classification systems are not sufficient in themselves and that comparable efforts must be directed towards a specification of the essential components of a reference or standard measure. Otherwise, at a time when public health authorities are keener than ever to obtain and use good epidemiological information (Murray & Lopez, 1996), psychiatric epidemiology will loose out (Bartlett & Coles, 1998). The World Health Organization is planing a world mental health survey that will be fielded in the year 2000. Data collection options for this survey are currently being discussed that include 'using clinical interviewer-based interviewing strategies that use clinical judgement to rate symptoms' (Kessler, 1999). If achievable, such a development would be a potentially significant advance on earlier efforts.

Before considering how greater clinical input into epidemiological assessment might be achieved, other less costly solutions should be discussed. One possibility is to make a statistical correction for the discrepancy between structured instruments and their semi-structured clinical equivalents. The standard instrument classifications can be used to adjust the indicator estimates. The precision of this procedure can be expressed by calculating, for example, confidence intervals. Our own group have recently reported initial work on a crosswalk between the CIS-R and SCAN in order to derive estimates of the prevalence (with confidence intervals) of psychiatric morbidity that would have been obtained had SCAN been used instead of the CIS-R in the British National Survey of Psychiatric Morbidity (Brugha *et al.* 1997). This will be published in more detail in the near future. Such an approach could significantly enhance comparison between surveys in different parts of the world.

Because structured interviews can be administered by lay survey interviewers, they have made it possible to collect epidemiological data in samples of 10000 or more (Eaton *et al.* 1984; Kessler *et al.* 1994; Jenkins *et al.* 1997). Epidemiological studies relying on clinically-based semi-structured measures such as the Present State Examination (Wing *et al.* 1974) and the Clinical Interview Schedule (CIS) (Goldberg *et al.* 1970) have been much smaller, rarely covering samples exceeding 1000 subjects (Jenkins *et al.* 1997). Larger samples than this have previously been thought impracticable.

The feasibility of training experienced survey interviewers to conduct semi-structured, clinical diagnostic interviewing across the range of neurotic and psychotic disorders has never been investigated systematically and reported. It could be argued that this would be an extension of the principle, endorsed by some structured interview developers (see above), that lay interviewers should have and can be given some clinical training.

Recently, we have worked with experienced survey interviewers from the Office for National Statistics (ONS), London. They underwent extended training in a Survey Form of SCAN (SCAN-SF). Eighty adults, including a majority of psychiatric in-patients were assessed by ONS interviewers and re-interviewed within a week by clinicians trained in SCAN (Brugha *et al.* 1999*a*). Sensitivity, specificity and concordance were calculated. Rater bias was also assessed, as this had been a problem in several community surveys in which lay interviewers conducted the PSE with householders (Sturt *et al.* 1981; Rodgers & Mann, 1986). Trainers found lay interviewers appeared to cope at least as well with psychotic as with neurotic disorders. Concordance for any disorder, for any specific psychotic disorder and for any specific neurotic disorder was good. Sensitivity and specificity were also good. There was no evidence of rater bias. These preliminary findings are very promising. However, before the SCAN-SF, administered by carefully trained lay interviewers, can be recommended in large scale surveys, further evaluations of its feasibility and reliability in the general population are needed.

CONCLUSION

A fundamental promise of the application of epidemiological methods to psychiatry is the collection of reliable information about the burden of psychiatric morbidity in whole populations. Two generations of attempts to achieve this have led to limited progress. The nature of psychiatric disorder is inherently difficult to define and therefore to assess reliably, particularly in non-referred samples. Future approaches are unlikely to be any more successful unless basic clinical judgements can be incorporated successfully in large-scale surveys by professional lay interviewers. Researchers need to debate constructively the core differences between structured and semi-structured approaches to assessment. Commensurate with the achievements of internationally standardized psychiatric classification, work towards a consensus specification of the essential components of a reference or standard measure is needed. Fully structured measures will continue to be used to provide mental health data in general health surveys with large sample sizes. However, a greater synthesis of clinical and structured methods is clearly what is needed in the future.

T. S. BRUGHA, P. E. BEBBINGTON AND R. JENKINS

REFERENCES

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, *4th edn*. American Psychiatric Association: Washington, DC.

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.

Bartlett, C. J. & Coles, E. C. (1998). 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.

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). 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, Australia.

Brugha, T. S., Nienhuis, F. J., Bagchi, D., Smith, J. & Meltzer, H. (1999*a*). 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.

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

Cooper, J. E., Copeland, J. R. M., Brown, G. W., Harris, T. & Gourlay, A. J. (1977). Further studies on interviewer training and inter-rater reliability of the Present State Examination (PSE). *Psychological Medicine* **7**, 517–523.

Eaton, W. W., Holzer, C. E., Von Korff, M., Anthony, J. C., Helzer, J. E., George, L., Burnam, M. A., Boyd, J. H., Kessler, L. G. & Locke, B. Z. (1984). The design of the Epidemiologic Catchment Area Surveys: the control and measurement of error. *Archives of General Psychiatry* **41**, 942–948.

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.

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). The national psychiatric morbidity surveys of Great Britain – strategy and methods. *Psychological Medicine* **27**, 765–774.

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., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U. & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**, 8–19.

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.

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.

Leighton, D. C., Harding, J. S., Macklin, D. B., MacMillan, A. M. & Leighton, A. H. (1963). *The Character of Danger: Psychiatric Symptoms in Selected Communities*. Basic Books: New York.

Lewis, G., Pelosi, A. J., Glover, E., Wilkinson, G., Stansfeld, S. A., Williams, P. & Shepherd, M. (1988). The development of a computerised assessment for minor psychiatric disorder. *Psychological Medicine* **18**, 737–745.

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.

Murray, C. J. L. & Lopez, A. D. (1996). *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020* (p. 43 Global burden of disease and injury). Harvard School of Public Health: Cambridge, MA.

Page, A. C. (1991). An assessment of structured diagnostic interviews for adult anxiety disorders. *International Review of Psychiatry* **3**, 265–278.

Pasamanick, B., Roberts, D. W. & Lemkau, P. V. (1956). A survey of mental disease in an urban population. *American Journal of Public Health* **47**, 923–929.

Pincus, A. L. & McQueen, L. E. (1997). US primary care training and the role of the DSM-IV Primary Care Version (DSM-IV-PC). In *Preventing Mental Illness: Mental Health Promotion in Primary Care* (ed. R. Jenkins and T. B. Üstün), pp. 229–252. John Wiley: Chichester.

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. *Archive of General Psychiatry* **55**, 109–115.

Robins, L. N. & Regier, D. A. (1991). *Psychiatric Disorders in America*. Free Press: New York.

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.

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.

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.

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. *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.

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.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., Joyce, P. R., Karam, E. G., Lee, C. K., Lellouch, J., Lepine, J. P., Newman, S. C., Rubio-Stipec, M., Wells, J. E., Wickramaratne, P. J., Wittchen, H. & Yeh, E. K. (1996). Cross-national epidemiology of major depression and bipolar disorder. *Journal of the American Medical Association* **276**, 293–299.

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. (1996). The WHO-Composite International Diagnostic Interview (CIDI): current status and the future. *European Psychiatry* **11**, 216s–217s. (abstract).

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.

Wittchen, H.-U., Kessler, R. C., Zhao, S. & Abelson, J. (1995). Reliability and clinical validity of UM-CIDI DSM-III-R generalized anxiety disorder. *Journal of Psychiatric Research* **29**, 95–110.

Wittchen, H.-U., Zhao, S., Abelson, J. M., Abelson, J. L. & Kessler, R. C. (1996). Reliability and procedural validity of UM-CIDI DSM-III-R phobic disorders. *Psychological Medicine* **26**, 1169–1177.

Wittchen, H.-U., Lachner, G., Wunderlich, U. & Pfister, H. (1998). Test–retest reliability of the computerised DSM-IV version of the Munich-Composite International Diagnostic Interview (M-CIDI). *Social Psychiatry and Psychiatric Epidemiology* **33**, 568–578.

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

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

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