292

Psychological Medicine, 1996, **26**, 1169–1177. Copyright © 1996 Cambridge University Press

Reliability and procedural validity of UM-CIDI DSM-III-R phobic disorders

H.-U. WITTCHEN,[1] S. ZHAO, J. M. ABELSON, J. L. ABELSON AND R. C. KESSLER

From the Max-Planck-Institut für Psychiatrie Klinisches Institut, Munich, Germany; and the Institute for Social Research, Department of Sociology and Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, USA

SYNOPSIS We evaluate the long-term test–retest reliability and procedural validity of phobia diagnoses in the UM-CIDI, the version of the Composite International Diagnostic Interview, used in the US National Co-morbidity Survey (NCS) and a number of other ongoing large-scale epidemiological surveys. Test–retest reliabilities of lifetime diagnoses of simple phobia, social phobia, and agoraphobia over a period between 16 and 34 months were $K = 0{\cdot}46$, 0·47, and 0·63, respectively. Concordances with the Structured Clinical Interview for DSM-III-R (SCID) were $K = 0{\cdot}45$, 0·62, and 0·63, respectively. Diagnostic discrepancies with the SCID were due to the UM-CIDI under-diagnosing. *Post hoc* analysis demonstrated that modification of UM-CIDI coding rules could dramatically improve cross-sectional procedural validity for both simple phobia ($K = 0{\cdot}57$) and social phobia ($K = 0{\cdot}95$). Based on these results, it seems likely that future modification of CIDI questions and coding rules could lead to substantial improvements in diagnostic validity.

INTRODUCTION

This report presents data on the test–retest reliability and procedural validity of the UM-CIDI (Wittchen *et al.* 1995), the version of the Composite International Diagnostic Interview (CIDI; WHO, 1990) developed at the University of Michigan (UM) for use in the US National Co-morbidity Survey (NCS; Kessler *et al.* 1994) and currently being used in a number of other epidemiological surveys around the world. The report focuses on the DSM-III-R diagnoses of simple phobia, social phobia, and agoraphobia (with or without panic). Phobias are among the most common psychiatric disorders in the population (Eaton *et al.* 1991; Kessler *et al.* 1994). They often begin early in life (Burke *et al.* 1991; Magee *et al.* 1996), cause significant role impairment (Magee *et al.* 1996), and are associated with the subsequent onset of depression, somatoform disorders, and substance use disorders (Wittchen *et al.* 1993; Magee *et al.* 1996). Some controversy continues to exist about the clinical significance of discriminating among simple, social, and agoraphobia (Rachman, 1985). However, the bulk of clinical and epidemiological data accumulated over the past decade argues for doing so based on evidence for differences in age of onset distributions (Magee *et al.* 1996), heritability (Kendler *et al.* 1992*a*) and treatment response (Foa & Kozak, 1985). This creates a need to develop methods to obtain reliable and valid diagnoses for the separate phobias.

The UM-CIDI is a fully structured diagnostic interview that generates diagnoses for the separate phobias as well as other psychiatric disorders using questions that are read word-for-word and response options that are recorded primarily in a Yes/No format. The greater standardization than in clinical interviews like the SADS (Fyer *et al.* 1985), SCID (Spitzer *et al.* 1990), or PSE (Wing *et al.* 1974) is needed because the UM-CIDI is designed to be used by trained interviewers who are not clinicians. The use of non-clinician interviewers is an important requirement for large and geographically dispersed epidemiological surveys, in which the use of clinical interviewers would not only pose daunting logistic problems but create prohibitive financial constraints on sample size.

[1] Address for correspondence: Professor Hans-Ulrich Wittchen, Max-Planck-Institut für Psychiatrie, Klinisches Institut, Kraeplinstrasse 10, D 80804 München, Germany.

Structured diagnostic interviews like the CIDI usually have higher test–retest reliability than clinical interviews (e.g. Williams *et al.* 1992; Wittchen, 1994). Consistent with this general finding, Wittchen (1994) found test–retest reliabilities with kappa (κ) values averaging 0·59 for simple phobia, 0·64 for social phobia and 0·68 for agoraphobia across a number of studies that administered fully structured diagnostic interviews in patient samples. These results compare favourably with the results of SCID reliability studies conducted in several patient samples, where test–retest reliabilities (κ) averaged 0·52 for simple phobia, 0·47 for social phobia and 0·43 for agoraphobia (Williams *et al.* 1992).

A concern can be raised that the good reliability of fully structured diagnostic interviews is obtained at the expense of clinical validity. Consistent with this possibility, studies of the procedural validity of fully structured assessments of DSM-III phobia have reported consistently poor levels of agreement with clinical re-interviews. For example, Helzer *et al.* (1985) reported relationships between diagnoses based on the fully structured Diagnostic Interview Schedule (DIS; Robins *et al.* 1981) and subsequent reinterviews by physicians who re-administered the DIS with a DSM-III diagnostic checklist of $\kappa = 0{\cdot}10$ for simple phobia, 0·15 for social phobia and 0·27 for agoraphobia in a community sample. Hwu *et al.* (1986), using the Chinese version of the DIS, reported a κ of 0·18 for undifferentiated phobic disorders, while Erdman *et al.* (1987) found κs between the DIS phobia diagnoses and clinical diagnoses ranging from a high of 0·22 for agoraphobia to $-0{\cdot}03$ for social phobia. However, there is evidence that agreement is higher when diagnoses are based on the more highly specified DSM-III-R criteria. Wittchen *et al.* (1994) found in a clinical reappraisal study of the CIDI using DSM-III-R criteria that agreement with a clinical re-interview had κ values of 0·3 for simple phobia, 0·6 for social phobia and 0·5 for agoraphobia. These agreement levels are quite similar to those found in test–retest studies of clinical interviews (Williams *et al.* 1992; DiNardo *et al.* 1993).

Unlike most previous methodological studies of the CIDI or its variants, the present report is based on a general population sample rather than on a treatment sample. UM-CIDI diagnoses are validated against blind clinical reinterviews based on the SCID (Spitzer *et al.* 1990). Furthermore, an attempt is made to explore reasons for discrepancies between UM-CIDI and SCID diagnoses by carrying out comparisons at the level of the diagnostic criterion and to evaluate the extent to which diagnostic concordance can be improved by modifying UM-CIDI rules for diagnostic classification.

METHOD

The sample

As described in more detail elsewhere (Kessler *et al.* 1994), the NCS was administered in face-to-face interviews to a nationally representative sample of 8098 respondents ages 15–54 from the non-institutionalized household population of the coterminous US between September, 1990 and February, 1992. The response rate was 82·4%. Clinical reappraisal interviews were administered in a series of separate diagnostic-specific NCS subsamples, with a minimum of 30 respondents in each subsample, at least 20 of whom were UM-CIDI/DSM-III-R cases defined without diagnostic hierarchy rules and at least 10 non-cases. A total of 40 respondents were re-interviewed in the simple phobia subsample (28 NCS cases and 12 non-cases), 37 in the social phobia subsample (23 NCS cases and 14 non-cases) and 34 in the agoraphobia subsample (23 NCS cases and 11 non-cases). The non-cases in each subsample consisted of respondents who endorsed the stem question for that phobia in the NCS, which asked about an unrealistically strong fear of certain objects or situations that invariably led either to extreme distress or avoidance, but failed to meet full diagnostic criteria. A more detailed discussion of the rationale for the design is presented elsewhere (Wittchen *et al.* 1995). The decision to select subthreshold controls instead of definite non-cases was based on the desire to provide a more sensitive evaluation of the extent to which the UM-CIDI correctly discriminates true cases from non-cases who are near the diagnostic threshold.

Measures

The UM-CIDI

As noted above, the UM-CIDI is a modified version of the Composite International Diagnostic Interview (CIDI: WHO, 1990). The CIDI was developed jointly by ADAMHA and WHO for purposes of standardizing psychiatric epidemiological research (Robins *et al.* 1988). Modifications included deleting some CIDI sections, modifying questions to clarify their meaning, rearranging question order to improve comprehension and flow and using visual checklists and review cards to simplify the complex cognitive tasks required of respondents. These changes were based on extensive pilot research (Kessler *et al.* 1996).

There were three UM-CIDI modifications in the assessment of phobias. First, the diagnostic sections were reordered so that the phobias were assessed prior to any other disorders. Secondly, respondents were presented with visual lists of potentially phobia objects and situations. Thirdly, an additional probe for avoidance was included to ask people who denied that their unreasonable fear persisted for months or years whether this was due to the fact that they always avoided the situation completely. The phobia stem questions and symptom questions were otherwise the same as in the CIDI.

The clinical reappraisal interviews

The clinical reappraisal interviews (available as part of the Internet appendix materials described in the acknowledgements along with more detailed discussion of data collection and coding procedures) were carried out by readministering the UM-CIDI phobia symptom questions followed by a modified version of the SCID section for the phobia under investigation. The re-interview began by telling respondents they would be asked some of the same questions as in the interview and that this was a test of the interview, not a test of their memory, so they should answer without trying to remember what they said to the other interviewer. They were then told that 'During the first interview, you reported (presentation of the stem question endorsed in the NCS interview)... I will be asking you some questions about this'.

This introduction was designed to minimize the possibility of the respondent attempting to remember his or her earlier answers and to force consistency in the report of the lifetime stem. The decision to force consistency in the stem question was based on past experience that re-interview respondents often deny stem questions endorsed in the baseline interview, leading clinical re-interviewers to declare that baseline diagnoses were invalid without clinically reappraising the symptoms reported in the structured interview (McLeod *et al.* 1990). Some previous studies have addressed this problem by carrying out a third interview that reviews discrepancies in reports in the first two interviews with the respondent in an effort to resolve inconsistencies (e.g. Manuzza *et al.* 1989; Williams *et al.* 1992). We rejected this option based on concerns about the difficulty of presenting inconsistencies to respondents in a way that did not make them defensive. We decided that a better strategy was to force consistency in the stem questions by means of the above introduction. Although, in theory, reappraisal interview respondents could have denied reporting an unrealistically strong fear in the NCS, none did so in the simple phobia re-interviews and only two each did so in the social phobia and agoraphobia re-interviews.

At the end of the introduction, the clinical interviewer readministered the UM-CIDI phobia section followed by an expanded version of the SCID focused on the diagnosis under investigation. The SCID skip rules were not used in order to guarantee that the supervisor and clinical reviewer (see below) could evaluate each criterion even if they disagreed with the interviewer concerning earlier criteria. There were a few cases in which it was not possible to recontact a respondent who reported SCID symptom data that the supervisors and clinical reviewer classified as inadequate. These respondents (5 of 40 for simple phobia, 6 of 37 for social phobia, and 6 of 34 for agoraphobia) were included in the evaluation of test–retest reliability but not in the evaluation of procedural validity.

Interviewer training and administration

Nine interviewers participated in the NCS clinical reappraisal study. One was a Ph.D. Clinical Psychologist, three were M.A. - level Clinical Psychologists, three held an MSW in Psychiatric Social Work, and two were B.A. - level Psychiatric Nurses. The interviews were all monitored by one of four clinical supervisors. One of these four was an M.D., one a Ph.D. in Clinical Psychology and two were M.A. - level Psychiatric Nurses. Final diagnoses were determined by an experienced clinical rater

(J. M. A.), based on a review of interviewer and supervisor materials supplemented by discussions with the interviewer and supervisor as well as a consulting psychologist (H. U. W.) and a consulting psychiatrist (J. L. A.) who are specialists in anxiety disorders.

The wide dispersion of the NCS over 172 counties in 34 states made it logistically impossible to carry out face-to-face clinical reappraisal interviews with a representative subsample of the entire NCS sample. As a result, our practical options for the reappraisal interviews were either to interview a local sample face-to-face or a national sample by telephone. We chose the latter option based on evidence that SCID interviews carried out over the telephone yield results similar to those carried out face-to-face (Kendler *et al.* 1992*b*; Sobin *et al.* 1993).

Analysis procedures

Agreement was analysed with the κ statistic (Cohen, 1960; Fleiss, 1981). Positive predictive value (PPV) and negative predictive value (NPV) were also calculated to disaggregate overall agreement into the components due to the percentage of NCS cases confirmed as cases (PPV) and due to the percentage of NCS non-cases confirmed as non-cases (NPV). Weighting adjustments for the over-sampling of NCS cases compared to non-cases were made prior to computing the agreement statistics. No adjustments for design effects were introduced into the calculation of standard errors of the estimates.

As described above, the non-cases in each re-interview subsample were confined to those who endorsed the stem question for that phobia. This provides a more sensitive evaluation than otherwise of the extent to which the UM-CIDI can correctly discriminate true cases from non-cases who are near the diagnostic threshold. However, it makes it impossible to estimate NPV, κ, sensitivity, or specificity. This problem was addressed by calculating an estimate of NPV for the subsample of respondents who endorsed the stem question (NPV1) and a separate upper-bound estimate of NPV based on the assumption that none of the NCS respondents who failed to endorse the stem question would have met criteria in the reappraisal interview (NPV2). The estimate of κ is also based on this same upper-bound assumption.

RESULTS

NCS distributions

Over half (52%) of NCS respondents endorsed the simple phobia stem question, over one-third (37%) endorsed the social phobia stem question, and over one-sixth (17·1%) endorsed the agoraphobia stem question. Full diagnostic criteria were met by about one-fifth of those who endorsed the stem question for simple phobia (11·5% of the total sample) and one-third for either social phobia (13·1% of the total sample) or agoraphobia (6·0% of the total sample).

Test–retest reliability and procedural validity: simple phobia

The test–retest reliability and procedural validity of the UM-DIDI diagnosis of simple phobia are presented in Part I of Table 1. Estimates of overall consistency are $\kappa = 0·46$ (reliability), 0·45 (prospective procedural validity), and 0·32 (cross-sectional procedural validity). Estimates of PPV are 0·57 for reliability and 0·82–0·83 for validity, meaning that while only 57% of NCS cases continued to report all the symptoms necessary to receive a diagnosis of simple phobia in the reinterview over 80% of the UM-CIDI cases in the NCS and re-interview were diagnosed as fulfilling all criteria for simple phobia by the clinical interviewers. Estimates of NPV1 are 0·83 for reliability and 0·44–0·64 for validity, meaning that while the vast majority of NCS respondents who endorsed the simple phobia stem question but not all other diagnostic criteria continued to be classified as non-cases in the UM-CIDI re-interview, the clinical interviewers were more likely to rate them as cases.

Inspection of the criterion-level measures of agreement in Part II of Table 1 shows that no one criterion is responsible for the low diagnostic reliability and validity. The criterion-level estimates of PPV are all very good and those of NPV1 are, for the most part, poor. Although more detailed analysis was unable to detect any way of recoding the UM-CIDI data to bring the diagnostic-level reliability and validity into the good range, this analysis led to the discovery that many of the false negatives responsible for the low NVP1 values were people who met criteria for three out of the four diagnostic criteria in the UM-CIDI. Based on this discovery we explored the implications of redefining UM-

Table 1. *Simple phobia: test–retest (TR) reliability and procedural validity of UM-CIDI Time 1 (V1C) and Time 2 (V2C) compared with clinical reappraisal with the SCID*

		PPV	(S.E.)	(*N*)	NPV2	(S.E.)	(*N*)	NPV1	(S.E.)	κ	(S.E.)
I. Diagnosis											
UM-CIDI	(TR)	0·57	(0·15)	(28)	0·83	(0·11)	(12)	0·92	(0·03)	0·46	(0·15)
	(V1C)	0·83	(0·11)	(24)	0·64	(0·15)	(11)	0·83	(0·04)	0·45	(0·12)
	(V2C)	0·82	(0·11)	(17)	0·44	(0·12)	(18)	0·75	(0.05)	0·32	(0·12)
Revised UM-CIDI	(TR)	0·96	(0·05)	(28)	0·58	(0·14)	(12)	0·82	(0·04)	0·58	(0·10)
	(V1C)	0·83	(0·09)	(24)	0·64	(0·15)	(11)	0·85	(0·04)	0·54	(0·11)
	(V2C)	0·79	(0·10)	(28)	0·71	(0·17)	(7)	0·88	(0·04)	0·57	(0·11)
II Diagnostic criteria											
A Persistence	(TR)	1·00	(0·00)	(27)	0·67	(0·27)	(3)	0·96	(0·03)	0·95	(0·03)
	(V1C)	1·00	(0·00)	(23)	0·50	(0·35)	(2)	0·94	(0·03)	0·93	(0·04)
	(V2C)	1·00	(0·00)	(34)	0·00	(0·00)	(1)	0·87	(0·04)	0·86	(0·05)
B Immediate anxiety response	(TR)	0·89	(0·06)	(28)	0·58	(0·14)	(12)	0·87	(0·04)	0·72	(0·07)
	(V1C)	0·97	(0·03)	(24)	0·36	(0·15)	(11)	0·80	(0·05)	0·69	(0·08)
	(V2C)	0·96	(0·04)	(23)	0·33	(0·14)	(12)	0·79	(0·05)	0·66	(0·08)
C Avoidance/endurance	(TR)	0·79	(0·11)	(30)	0·78	(0·14)	(10)	0·87	(0·04)	0·53	(0·12)
	(V1C)	0·87	(0·09)	(26)	0·78	(0·14)	(9)	0·90	(0·03)	0·64	(0·10)
	(V2C)	0·95	(0·06)	(24)	0·36	(0·13)	(11)	0·72	(0·05)	0·40	(0·10)
D Interference/distress	(TR)	0·96	(0·05)	(28)	0·42	(0·15)	(12)	0·73	(0·05)	0·37	(0·11)
	(V1C)	0·88	(0·10)	(24)	0·55	(0·15)	(11)	0·79	(0·04)	0·41	(0·11)
	(V2C)	0·88	(0·10)	(29)	0·67	(0·19)	(6)	0·85	(0·04)	0·50	(0·12)

(TR) = Test–Retest.
(V1C) = Validity comparison between the UM-CIDI in the baseline NCS and the SCID.
(V2C) = Validity comparison between the UM-CIDI in the reappraisal interview and the SCID.

CIDI cases as those who met at least three of the four diagnostic criteria. As shown in Part I, this revised coding scheme led to an improvement in κ for both reliability (0·58) and validity (0·54–0·57).

Test–retest reliability and procedural validity: social phobia

The test–retest reliability and procedural validity of the UM-CIDI diagnosis of social phobia are presented in Part I of Table 2. Estimates are $\kappa =$ 0·47 (reliability), 0·62 (prospective procedural validity), and 0·54 (cross-sectional procedural validity). Estimates of PPV are quite good (0·74 for reliability and 0·91–1·0 for validity), while estimates of NPV1 are poor (0·50 for reliability and 0·44–0·50 for validity). This means that the UM-CIDI under-diagnosed social phobia in the NCS.

The criterion-level measures of agreement in Part II of Table 2 show that PPV is good for all criteria while NPV1 is poor for all criteria. As in the case of simple phobia, many of respondents responsible for the low NVP1 met criteria for three out of the four diagnostic criteria in the UM-CIDI. As shown in Part I of Table 2, redefining UM-CIDI cases as those who met three or more of the UM-CIDI criteria led to a dramatic improvement in the estimated κ for both reliability (0·58) and validity (0·68–0·95) of this diagnosis.

Test–retest reliability and procedural validity: agoraphobia

The test–retest reliability and procedural validity of the UM-CIDI diagnosis of agoraphobia are presented in Part I of Table 3. Estimates are considerably higher than for the other phobias: $\kappa = 0{\cdot}63$ (reliability), 0·63 (prospective procedural validity) and 0·79 (cross-sectional procedural validity). Estimates of PPV are lower than for the other phobias (0·57 for reliability and 0·64–0·92 for validity), while estimates of NPV1 are a good deal higher (0·63 for reliability and 0·63–0·79 for validity). This combination of low PPV and high NPV1 for a disorder such as agoraphobia, in which the number of respondents who endorsed the stem question but did not meet full diagnostic criteria (11·1 %) is considerably larger than the number who met full diagnostic criteria (6·0 %), means that the UM-CIDI over-diagnosed the prevalence of agoraphobia in the NCS.

Analysis of criterion-level concordance failed to detect any way of recoding the UM-CIDI data that would both maintain the good di-

Table 2. *Social phobia: test–retest (TR) reliability and procedural validity of UM-CIDI Time* 1 *(V*1*C) and Time* 2 *(V*2*C) compared with clinical reappraisal with the SCID*

		PPV	(S.E.)	(*N*)	NPV2	(S.E.)	(*N*)	NPV1	(S.E.)	*κ*	(S.E.)
I. Diagnosis											
UM-CIDI	(TR)	0·74	(0·12)	(23)	0·50	(0·13)	(14)	0·86	(0·04)	0·47	(0·12)
	(V1C)	1·00	(0·00)	(19)	0·50	(0·14)	(12)	0·86	(0·04)	0·62	(0·10)
	(V2C)	0·91	(0·08)	(22)	0·44	(0·17)	(9)	0·85	(0·04)	0·54	(0·11)
Revised UM-CIDI	(TR)	0·92	(0·07)	(25)	0·33	(0·14)	(12)	0·84	(0·04)	0·58	(0·10)
	(V1C)	0·95	(0·05)	(22)	0·50	(0·16)	(9)	0·88	(0·04)	0·68	(0·09)
	(V2C)	0·93	(0·06)	(27)	1·00	(0·00)	(4)	1·00	(0·00)	0·95	(0·04)
II Diagnostic criteria											
A/F Persistence	(TR)	0·83	(0·07)	(33)	0·25	(0·22)	(4)	0·92	(0·03)	0·75	(0·07)
	(V1C)	0·81	(0·07)	(28)	0·67	(0·27)	(3)	0·97	(0·02)	0·80	(0·07)
	(V2C)	0·76	(0·08)	(27)	0·00	(0·00)	(4)	0·95	(0·03)	0·73	(0·08)
C Immediate anxiety response	(TR)	0·87	(0·09)	(23)	0·07	(0·07)	(14)	0·74	(0·05)	0·37	(0·11)
	(V1C)	1·00	(0·11)	(19)	0·08	(0·08)	(12)	0·75	(0·05)	0·44	(0·11)
	(V2C)	1·00	(0·00)	(29)	0·00	(0·00)	(2)	0·72	(0·05)	0·41	(0·11)
D Avoidance/endurance	(TR)	0·95	(0·04)	(29)	0·38	(0·17)	(8)	0·90	(0·03)	0·79	(0·07)
	(V1C)	1·00	(0·00)	(25)	0·67	(0·19)	(6)	0·95	(0·03)	0·90	(0·05)
	(V2C)	0·90	(0·06)	(28)	0·67	(0·27)	(3)	0·95	(0·03)	0·83	(0·06)
E Interference/distress	(TR)	0·84	(0·10)	(24)	0·31	(0·13)	(13)	0·81	(0·04)	0·46	(0·11)
	(V1C)	0·94	(0·06)	(20)	0·45	(0·15)	(11)	0·85	(0·04)	0·59	(0·11)
	(V2C)	0·91	(0·08)	(25)	0·67	(0·19)	(6)	0·91	(0·03)	0·69	(0·10)

(TR) = Test–Retest.
(V1C) = Validity comparison between the UM-CIDI in the baseline NCS and the SCID.
(V2C) = Validity comparison between the UM-CIDI in the reappraisal interview and the SCID.

agnostic validity of the original coding scheme and reduce the over-estimation of agoraphobia prevalence based on this scheme. In addition, as shown in Part II of Table 3, it is quite common for UM-CIDI non-cases to be classified as meeting either Criterion A or Criterion B, but not both, in the SCID, making it difficult to modify the UM-CIDI coding scheme for agoraphobia in a fashion similar to the modifications developed for a simple phobia and social phobia without dramatically reducing NPV1.

DISCUSSION

Design considerations

Several features of the research design limit our ability to make direct comparisons with these prior studies. First, this study was carried out in a general population sample, while most previous studies of the CIDI were conducted in clinical samples. Secondly, we focused on long-term reliability and validity, while most previous studies examined primarily short-term (usually with a time interval of 1 to 3 days between interviews) agreement. Thirdly, the reappraisal interviews with NCS non-cases focused on respondents who endorsed the stem question for the disorder in the NCS, providing a much more sensitive evaluation of whether the UM-CIDI can correctly discriminate true cases from non-cases near the diagnostic threshold than previous studies. The fact that we forced consistency in the report of stem questions presumably led to an increase in estimated reliability, but the focus on non-cases who endorsed the stem question made it more difficult than in previous studies to document consistency in reports.

Test–retest reliability

In light of these special design features, most of which make it considerably harder than in previous studies to obtain high agreement coefficients, our findings document acceptable *κ* coefficients for the test–retest reliability of all three types of phobias. These findings compare favourably to the joint analysis of three independent short-term (1–3 days) reliability studies by Semler *et al.* (1988), Wacker (1991) and Wittchen (1994). Furthermore, criterion-level analyses showed that the vast majority of the test–retest diagnostic discrepancies were due to respondents being inconsistent over time in only one diagnostic criterion.

The inconsistencies between the NCS and re-interview UM-CIDI reports of simple phobia can be traced largely to questions that assess the DSM-III-R interference, distress, avoidance, and endurance criteria. We can speculate that

Table 3. *Agoraphobia: test–retest (TR) reliability and procedural validity of UM-CIDI Time* 1 *(V1C) and Time* 2 *(V2C) compared with clinical reappraisal with the SCID*

		PPV	(S.E.)	(*N*)	NPV2	(S.E.)	(*N*)	NPV1	(S.E.)	κ	(S.E.)
I. Diagnosis											
	(TR)	0·57	(0·20)	(23)	0·91	(0·09)	(11)	0·99	(0·01)	0·63	(0·19)
	(V1C)	0·64	(0·20)	(22)	0·83	(0·15)	(6)	0·98	(0·01)	0·63	(0·18)
	(V2C)	0·92	(0·11)	(13)	0·80	(0·10)	(15)	0·98	(0·02)	0·79	(0·13)
II Diagnostic criteria											
A Fear	(TR)	0·61	(0·20)	(32)	0·50	(0·22)	(4)	0·94	(0·02)	0·44	(0·19)
	(V1C)	0·68	(0·19)	(28)	—	(—)	(0)	—	(—)	0·80	(0·14)
	(V2C)	0·92	(0·11)	(14)	0·64	(0·13)	(14)	0·96	(0·02)	0·69	(0·14)
A Consequences	(TR)	0·68	(0·12)	(23)	0·27	(0·13)	(11)	0·97	(0·02)	0·70	(0·11)
	(V1C)	0·58	(0·13)	(22)	0·67	(0·19)	(6)	0·99	(0·01)	0·66	(0·12)
	(V2C)	0·58	(0·13)	(25)	0·67	(0·27)	(3)	0·99	(0·01)	0·67	(0·12)

(TR) = Test–Retest.
(V1C) = Validity comparison between the UM-CIDI in the baseline NCS and the SCID.
(V2C) = Validity comparison between the UM-CIDI in the reappraisal interview and the SCID.

inconsistencies in respondent answers to these lifetime questions over time are influenced by the current severity of the disorder and its current psychosocial consequences. For social phobia, Criterion C of DSM-III-R, requiring at some stage of the disorder an immediate anxiety reaction upon exposure, was found to be the most common source of inconsistency in addition to Criterion E (significant interference). It should also be noted that we did not find any drop-off in the number of cases from Time 1 to Time 2, a problem that has plagued previous reliability studies (Helzer *et al.* 1977; Semler *et al.*, 1988).

Procedural validity

The clinical reappraisal of the baseline UM – CIDI with the SCID confirmed 83 % of all NCS cases with simple phobia, 100 % of cases with social phobia and 64 % of cases with agoraphobia. The positive predictive values for the UM-CIDI with the SCID were all excellent: 82 % for simple phobia, 91 % for social phobia and 92 % for agoraphobia, meaning that respondents classified as cases by the UM-CIDI would very likely have been classified as cases based on a clinical interview. For agoraphobia, these good PPVs were matched by high negative predictive values, ranging from 80–83 %, resulting in good κ values for both prospective (0·63) and cross-sectional (0·79) procedural validity. The situation was considerably worse, though, for simple and social phobias, where NPVs ranged from 75 % (simple phobia, cross-sectional) to 86 % (social phobia, lagged). Given that roughly half of all NCS respondents endorsed the stem question for simple phobia and more than one-third endorsed the stem question for social phobia, these comparatively low NPV1 values mean that there are a substantial number of SCID cases who were classified as non-cases in the UM-CIDI, resulting in low κ values of 0·32 (cross-sectional) and 0·45 (longitudinal) for simple phobia and 0·54 (cross-sectional) and 0·62 (longitudinal) for social phobia. The higher κ values for social phobia than simple phobia are a result of the slightly higher PPVs and the fact that a lower proportion of NCS respondents endorsed the stem question. Assuming that our clinical reappraisal with the SCID is an appropriate gold-standard against which we can evaluate the quality of the CIDI, these findings suggest the UM-CIDI question together with its associated diagnostic algorithms are too strict, resulting in a substantial under-estimation of the true prevalence of simple and social phobias.

If we were to accept the SCID diagnoses as a gold standard and readjust the NCS prevalence estimates based on the validity findings, the estimated prevalence of agoraphobia would not differ meaningfully from the estimate in the NCS. The estimated prevalences for simple phobia and social phobia, however, would increase markedly – from 11·5 % in the NCS to 24·1 % for simple phobia and from 13·1 % in the NCS to 25·1 % for social phobia. However, these estimates all have large standard errors due to the small size of the reappraisal interview samples and they are based on the implausible assumption that the SCID diagnoses are perfectly accurate. We know from methodological studies on the SCID (Wittchen *et al.* 1990; Williams *et al.* 1992) that the latter is not the case. Indeed, these methodological studies show that the SCID diagnoses of simple phobia and social phobia have only modest test–retest reliability in clinical samples. Therefore, while it is clear from the validation results that the UM-CIDI under-estimates the prevalences of simple phobia and social phobia, there is considerable uncertainty as to the magnitudes of these effects.

More detailed analyses of criterion-level validities were carried out to develop some understanding of the UM-CIDI questions that play the most important part in these under-estimates. Based on these results, it appears that the least validly assessed questions in the UM-CIDI are those designed to assess the immediate anxiety reaction upon exposure criterion for social phobia, the avoidance/endurance criterion for simple phobia, and the interference/distress criterion for both simple and social phobia. These conclusions need replication in an independent study due to the fact that we only studied a small number of cases here. If these results are replicated, it would imply that improvement in these UM-CIDI diagnoses could best be achieved by changing these particular questions.

Prior to doing this, we developed a revised coding scheme that addressed the UM-CIDI under-diagnosis problem by using a more generous diagnostic decision rule that endorsement of any three of the four diagnostic criteria for simple phobia or social phobia resulted in a UM-CIDI diagnosis. This revision led to a substantial improvement in the validity of the assessment of both simple and social phobia. Based on these results, it appears likely that modification of the CIDI and more systematic empirical analysis of CIDI diagnosis classification rules would result in even further improvements in instrument validity.

The National Co-morbidity Survey (NCS) is a collaborative epidemiological investigation of the prevalence, causes and consequences of psychiatric morbidity and co-morbidity in the United States supported by the National Institute of Mental Health (Grants R01 MH46376 and R01 MH49098), National Institute of Drug Abuse (through MH46376), and the W. T. Grant Foundation (Grant 90135190). Collaborating NCS sites and investigators are: The Addiction Research Foundation (Robin Room), Duke University (Dan Blazer, Marvin Swartz), Johns Hopkins University (James Anthony, William Eaton, Philip Leaf), the Max Planck Institute of Psychiatry – Clinical Institute (Hans-Ulrich Wittchen), the Medical College of Virginia (Kenneth Kendler), the University of Michigan (Lloyd Johnston, Ronald Kessler), New York University (Patrick Shrout), SUNY Stony Brook (Evelyn Bromet), The University of Toronto (R. Jay Turner) and Washington University (Linda Cottler). Preparation of this report was also supported by a Research Scientist Award to Kessler (Grant K05 MH00507). We thank Monica Greene for training the clinical interviewers to use the SCID, Holli Bertram, David Schroat, and Beth Sherman for carrying out the SCID interviews, Michael First for helpful consultation and Cindy Huntimer for her help in all phases of the study. Text of NCS publications and study documentation can be obtained from the Internet by : FTP 141.211.207.206; NAME = anonymous; PASSWORD = (your Internet e-mail address); read the INSTRUCTIONS to download documents, including appendix materials for the current report.

REFERENCES

Burke, K. C., Burke, J. D., Rae, D. S. & Regier, D. A. (1991). Comparing age at onset of major depression and other psychiatric disorders by birth cohorts in five US community populations. *Archives of General Psychiatry* **48**, 789–795.

Cohen, J. (1960). A coefficient of agreement for nominal data. *Educational and Psychological Measurement* **20**, 37–46.

DiNardo, P. A., Moras, K., Barlow, D. H., Rapee, R. M. & Brown, T. A. (1993). Reliability of DSM-III-R anxiety disorder categories. Using the anxiety disorders interview schedule-revised (ADIS-R). *Archives of General Psychiatry* **50**, 251–256.

Eaton, W. W., Dryman, A. & Weissman, M. M. (1991). Panic and phobia: the diagnosis of panic disorder and phobic disorder. In *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study* (ed. L. N. Robins and D. A. Regier), pp. 155–179. The Free Press: New York.

Erdman, H. P., Klein, M. H., Greist, J. H., Bass, S. M., Bires, J. K. & Machtinger, P. E. (1987). A comparison of the diagnostic interview schedule and clinical diagnosis. *American Journal of Psychiatry* **144**, 1477–1480.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edn. John Wiley and Sons: New York.

Foa, E. B. & Kozak, M. J. (1985). Treatment of anxiety disorders: implications for psychopathology. In *Anxiety and the Anxiety Disorders* (ed. A. H. Tuma and J. D. Maser), pp. 421–452. Erlbaum: Hillsdale.

Fyer, A. J., Endicott, J., Mannuzza, S. & Klein, D. F. (1985). *Schedule for Affective Disorders and Schizophrenia – Lifetime Versions (SADS-LA)*. New York State Psychiatric Institute: New York.

Helzer, J. E., Clayton, P. J., Pambakian R., Reich, T., Woodruff, R. A., Jr. & Reveley, M. A. (1977). Reliability of psychiatric diagnosis. II. The test – retest reliability of diagnostic classification. *Archives of General Psychiatry* **34**, 136–141.

Helzer, J. E., Robins, L. N., McEvoy, L. T. & Spitznagel, E. (1985). A comparison of clinical and diagnostic interview schedule diagnoses. *Archives of General Psychiatry* **42**, 657–666.

Hwu, H.-G., Yeh, E.-K. & Chang, L.-Y. (1986). Chinese diagnostic interview schedule. I. Agreement with psychiatrist's diagnosis. *Acta Psychiatrica Scandinavica* **73**, 225–233.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1992*a*). The genetic epidemiology of phobias in women: the inter-relationship of agoraphobia, social phobia, situational phobia and simple phobia. *Archives of General Psychiatry* **49**, 273–281.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. (1992*b*). A population-based twin study of major depression in women. The impact of varying definitions of illness. *Archives of General Psychiatry* **49**, 257–266.

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., Mroczek, D. K. & Belli, R. F. (1996). Retrospective adult assessment of childhood psychopathology. In *Assessment in Child Psychopathology* (ed. D. Shaffer and J. Richters). Guilford Press: New York. (In the press.)

McLeod, J. D., Turnball, 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.

Magee, W. J., Eaton, W. W., Wittchen, H.-U., McGonagle, K. A. & Kessler, R. C. (1996). Agoraphobia, simple phobia, and social phobia in the National Comorbidity Survey. *Archives of General Psychiatry* **53**, 159–168.

Mannuzza, S., Fyer, A. J., Martin, L. Y., Gallops, M. S., Endicott, J., Gorman, J., Liebowitz, M. R. & Klein, D. F. (1989). Reliability of anxiety assessment. I. Diagnostic agreement. *Archives of General Psychiatry* **46**, 1093–1101.

Rachman, S. (1985). The treatment of anxiety disorders: a critique of the implications for psychopathology. In *Anxiety and the Anxiety Disorders* (ed. A. H. Tuma and J. D. Maser), pp. 453–462. Erlbaum: Hillsdale.

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, L. H. (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.

Semler, G., Wittchen, H.-U. & Zaudig, M. (1988). The test-retest reliability of the german version of the Composite International Diagnostic Interview on RDC diagnoses and symptom level. In *International Classification in Psychiatry: Unity and Diversity* (ed. J. Mezzich and M. Cranach von), pp. 221–234. Cambridge University Press: Cambridge.

Sobin, C., Weissman, M. M., Goldstein, R. B., Adams, P., Wickramaratne, P., Warner, V. & Lish, J. D. (1993). Diagnostic interviewing for family studies: Comparing telephone and face-to-face methods for the diagnosis of lifetime psychiatric disorders. *Psychiatric Genetics* **3**, 227–233.

Spitzer, R. L., Williams, J. B. W., Gibbons, M. & First, M. B. (1990). *The Structured Clinical Interview for DSM-III-R, Patient Edition* (SCID-P, Version 1.0). American Psychiatric Press: Washington, DC.

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

Wacker, H. R. (1991). *Angst und Depression: Deskriptoren, Prädiktoren: Eine Querschnitts und Verlaufsuntersuchung.* Schweizer Nationalfonds. Schlußbericht. Gesuch-Nr. 32–9373 (3.995–0.87): Bern.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Rounsaville, B. & Wittchen, H.-U. (1992). The structured clinical interview for DSM-III-R (SCID) II: Multi-site test–retest reliability. *Archives of General Psychiatry* **49**, 630–636.

Wing, J. K., Cooper, J. E. & Sartorius, N. (1974). *Measurement and Classification of Psychiatric Symptoms: An Instruction Manual for the PSE and CATEGO Program.* 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., Schramm, E., Zaudig, M., Spengler, P., Rummler, R. & Mombour, W. (1990). *SKID: Strukturiertes Klinisches Interview für DSM-III-R: Manual; Version 2.0.* Beltz Verlag: Weinheim.

Wittchen, H.-U., Essau, C. A., Rief, W. & Fichter, M. M. (1993). Assessment of somatoform disorders and comorbidity pattern with the CIDI-findings in psychosomatic inpatients. *International Journal of Methods in Psychiatric Research* **3**, 87–100.

Wittchen, H.-U., Zhao, S., Kessler, R. & Eaton, W. W. (1994). DSM-III-R generalized anxiety disorder in the National Comorbidity Survey. *Archives of General Psychiatry* **51**, 355–364.

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

World Health Organization (1990). *Composite International Diagnostic Interview* (CIDI): (a) CIDI-interview (version 1.0); (b) CIDI-user manual; (c) CIDI-training manual; (d) CIDI-computer programs. World Health Organization: Geneva.