Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS)

RONALD C. KESSLER, Department of Health Care Policy, Harvard Medical School, Boston MA, USA
HANS-ULRICH WITTCHEN, Max Planck Institute of Psychiatry, Munich, Germany
JAMIE M. ABELSON, KATHERINE MCGONAGLE, NORBERT SCHWARZ, Institute for Social Research, University of Michigan, Ann Arbor MI, USA
KENNETH S. KENDLER, Department of Psychiatry and Human Genetics, Medical College of Virginia, Richmond, VA, USA
BÄRBEL KNÄUPER, Department of Psychology, Free University of Berlin, Berlin, Germany
SHANYANG ZHAO, Department of Sociology, Temple University, Philadelphia, Pennsylvania, USA

ABSTRACT *This paper reports the results of methodological studies carried out in conjunction with the US National Comorbidity Survey (NCS) to evaluate Version 1.0 of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). These studies relied on recent survey data collection methodology literature to investigate problems regarding question comprehension, instruction comprehension, respondent motivation to report accurately, and regarding the limits of respondent ability to report accurately. Insights and strategies developed by survey methodologists were used to modify the CIDI in an effort to address these problems. The paper describes these strategies and methodological studies that evaluated their effects, including a clinical reappraisal study and a field experiment that evaluated the impact of question modifications on prevalence estimates. The paper closes with a discussion of remaining methodological problems with the CIDI and potentially useful future studies that might be able to develop solutions to these problems.*

Introduction

This paper presents results of methodological studies carried out in conjunction with the US National Comorbidity Survey (NCS) (Kessler et al. 1994). These studies focused on the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI) (WHO 1990), the diagnostic interview used in the NCS. Results are presented both of NCS pretests that led to CID modifications and of subsequent methodological studies aimed at evaluating the effects of these modifications.

The NCS was a large nationally representative survey of the US household population designed to estimate the prevalence and correlates of DSM-III-R psychiatric disorders. It was necessary to use a fully structured instrument like the CIDI in the NCS because of the large sample size, the enormous geographic dispersion of the sample, and the prohibitive costs and logistic complications of the study. The NCS was the first large-scale general population survey to administer the CIDI in the United States. Although WHO CIDI Field Trials carried out prior to the NCS documented good performance of the instrument (Wittchen 1994), the field trials were conducted largely in clinical samples and administered in clinical settings. As a result, we considered it very important to carry out pretests before using the instrument in a community sample.

As described more fully below, the NCS pretests were guided by the literature on survey data collection methodology (e.g. Bradburn et al. 1979; Moss and

Goldstein 1979; Biderman 1980; Cannell et al. 1981; Jabine et al. 1984; Tanur 1992; Schwarz and Sudman 1992; 1994; 1996). Sudman, Bradburn and Schwarz (1996) provide a comprehensive review of these developments. This literature provided a number of important insights and strategies that were used to make modifications to the CIDI in sections of the instrument where the pretests documented comprehension problems. The paper closes with a discussion of remaining methodological problems with the CIDI and potentially useful future studies that might be able to develop solutions to these problems.

Methods and procedures

Six separate methodological studies of the CIDI were carried out in conjunction with the NCS. The methods and procedures used in each of these studies are described in this section of the paper. We then turn to a presentation of study-specific results. The first three of these six studies were non-experimental pretests based on convenience samples that were administered by experienced survey interviewers trained in the CIDI. Modifications to the CIDI were made in the NCS based on the results of these pretests. The other three were carried out subsequent to the NCS in an effort to evaluate the effects of the CIDI modifications. The first was a field experiment that evaluated the impact of the CIDI modifications on prevalence estimates. The second was a *post hoc* comparison of symptom and diagnostic prevalence estimates in the NCS versus the Epidemiologic Catchment Area Study (Robins and Regier 1991). The third was a clinical reappraisal study of the CIDI diagnostic classifications in the NCS based on blind reinterviews carried out by clinicians using the Structured Clinical Interview for DSM-III-R (SCID; Spitzer et al. 1992).

The NCS pretests

The first pretest was carried out in a quota sample of 30 volunteers in the age range 15–54, recruited by household screening, in the Detroit Metropolitan Area. Roughly equal numbers of men (n = 13) and women (n = 17) were interviewed, half of them selected from a working-class neighbourhood and the other half from a middle-class neighbourhood. Interviews were carried out face to face in the homes of respondents and were followed by a telephone-debriefing interview. A $25 financial incentive was offered for participation. As described below, the main focus of this first pretest was on discovering comprehension problems in the CIDI.

The second pretest was carried out in a quota sample of 50 volunteers (25 men and 25 women) in the age range 18–54 residing in Ann Arbor, Michigan, who were recruited from advertisements. Half of these respondents had no college education and the other half were college graduates. Interviews were carried out in the homes of respondents. A $25 financial incentive was offered for participation. This second pretest had two purposes. The first was to evaluate the effects of question wording changes made to correct the comprehension problems discovered in the first pretest. The second purpose was to evaluate a series of strategies aimed at motivating serious and accurate memory search and reporting.

The third pretest was carried out in a quota sample of 50 volunteers (22 men and 28 women) in the age range 15–54, sampled and interviewed in their homes using the same sampling frame and recruitment procedures as in the first pretest as well as in a small convenience sample of patients interviewed at a local community mental health centre. A $25 financial incentive was offered for participation. The purpose of this pretest was to obtain timing estimates and to make small revisions to the near-final version of the full survey instrument to be used in the NCS. The instrument tested included both the modified CIDI and detailed batteries of questions on risk factors and social consequences of psychiatric disorders developed for the survey. Unlike the earlier pretests, no debriefing or special methodological probing was used in this third pretest as the main purpose was to simulate production interviewing as closely as possible. No results from this pretest are therefore presented in this paper.

The experimental comparison of the original and modified versions of the CIDI

The remaining three methodological studies consisted of evaluations of the final instrument used in the NCS. The first of these three studies was an experimental evaluation in which a nationally representative quota sample of respondents in the age range 18–54 residing in households with telephones was given the introduction and the anxiety and mood disorders sections from either the original WHO CIDI Version 1.0 (n = 105) or the modified version of the CIDI included in the NCS (n = 106). These interviews were carried out by experienced telephone interviewers who were trained in either the original CIDI or the modified CIDI. The purpose was to make a global evaluation of the impact

of our modifications on lifetime prevalence estimates of DSM-III-R disorders.

Sampling for this pretest began with a geographically stratified and nationally representative sample of listed telephone numbers. The 'most recent birthday method' was used to select a random respondent of a prespecified sex from those at home at the time of contact. Contact attempts were made in the evenings and on weekends to guarantee that employed people had an adequate chance of being selected. The proportion of males and females was varied over the course of the study in order to guarantee a distribution of men and women that was as equal as possible. The target sample size was 100 interviews in each condition and the study was terminated at the end of the week in which this quota was achieved. No financial incentive was offered for participation. A 62% cooperation rate was obtained, including 51 men and 54 women randomly assigned to the WHO-CIDI condition and 52 men and 54 women assigned to the modified CIDI condition.

The post hoc comparison of the NCS with the ECA

The second methodological study after the pretests was a study of similarities and differences in patterns of symptom reporting and estimated diagnostic prevalences for disorders that were assessed in a comparable way in the NCS and the Epidemiologic Catchment Area (ECA) study (Robins and Regier 1991). As the ECA was based on DSM-III criteria and the NCS on DSM-III-R criteria, comparisons were made by recoding the NCS data to DSM-III criteria. As most of the results of the comparative NCS-ECA study have been reported previously (Regier et al. 1997), only highlights of this investigation relevant to other aspects of NCS methodological studies are presented here.

The clinical reappraisal study

The final methodological study was a clinical reappraisal study in which small diagnosis-specific random subsamples of NCS respondents who endorsed stem questions for particular diagnoses were recontacted by clinical interviewers who administered an expanded version of the Structured Clinical Interview for DSM-III-R (SCID) (Spitzer et al. 1992). Stem questions are the first questions in each diagnostic section, which ask the subject whether he or she ever experienced core symptoms of the syndrome under investigation. For example, the first question in the panic section of the CIDI asks the respondents if they ever had sudden spells or attacks of feeling very frightened or anxious. Only the subsample of respondents that endorses such stem questions is administered the remaining questions in a diagnostic section.

Aspects of the clinical reappraisal study have been reported previously (Blazer et al. 1994; Kessler et al. 1995; Warner et al. 1995; Wittchen et al. 1995, 1996; Kendler et al. 1997; Kessler, Crum et al. 1997). However, this is the first time the full set of diagnosis-level results has been presented. The clinical reappraisal study was carried out in 10 separate disorder-specific subsamples of NCS respondents rather than in one large subsample in order to minimize the burden of the reinterview on individual respondents. Clinical interviewers were blind to the original CIDI diagnoses in the NCS. Reinterviews took place between 13 and 33 months after the initial NCS interview.

Participants included subsamples of at least 30 respondents (at least 20 of whom fulfilled CIDI/DSM-III-R criteria and at least 10 non-cases) for each diagnosis included in the NCS. This sample size decision was based on the fact that it yields a standard error between 0.10 and 0.15 for κ in the range 0.60–0.80. The decision to select controls who endorsed the stem questions was based on the desire to provide a more sensitive evaluation of the extent to which the modified CIDI can correctly discriminate true cases from non-cases who are near the diagnostic threshold than if the controls had included respondents who failed to endorse the stem questions.

The interviewers in the clinical reappraisal study included a physician, clinical psychologists, psychiatric social workers, and psychiatric nurses. The interviewers had between two and 19 years of clinical experience. Each interviewer was trained on a single diagnostic section of the CIDI and SCID at a time and worked on that section until all interviews were completed. Interviewers were recruited and assigned to work on particular sections based on their expertise. For example, interviewers who worked on the substance disorders sections were recruited from the clinical staff of the alcohol and drug abuse service at the University of Michigan Medical Center. Clinical diagnoses were based on consensus ratings between the interviewer and the supervisor (JMA). Discrepancies were resolved in consensus conferences with senior clinical consultants. The clinical reappraisal interviews were carried out by readministering the relevant modified CIDI section followed by a modified version of the SCID for that same diagnosis. The reinterview

began by introducing the task in the following way:

> I am going to be briefly asking you some of the same questions you were asked on [date of initial interview]. This is a test of the interview and not a test of your memory, so answer the questions as completely as you can without trying to remember what you said to the other interviewer. During the first interview, you reported that [diagnostic stem question]. I will be asking you some questions about this.

At the end of the introduction, the clinical interviewer administered the modified CIDI section, excluding the stem question, followed by a modified version of the SCID for the same diagnosis. The first SCID modification was that we omitted the introductory overview section conventionally used, in which the interviewer asks about the patient's demographic characteristics, past and current psychopathology, role functioning, and life stressors. Second, the SCID 'skip rules' to omit further questions after failing a required criterion were not used. This modification was introduced in order to guarantee that the clinical supervisor and clinical consultants could evaluate each criterion, even if they disagreed with the interviewer about early criteria, as well as to allow an assessment of agreement and disagreement between the modified CIDI and the SCID for each criterion. Third, the interviewer was required to ask the mandatory SCID criterion questions without assuming an answer even if the criterion information was already obtained from a similar CIDI retest question. Fourth, the required SCID symptom and probe questions were supplemented with additional probes in an attempt to guarantee that the clinical interviewer evaluated each criterion fully and consistently. It is important to note that the inclusion of these additional probes is consistent with the SCID administration guidelines, which allows the interviewer to use additional probes whenever the answers to the initial standard SCID questions do not allow a symptom to be rated with certainty.

The introduction was designed to minimise the possibility of the respondent consciously attempting to remember his or her earlier answers, something that could lead to an artificially high estimate of test–retest reliability, while simultaneously forcing consistency in the report of the lifetime stem question. The decision to force consistency in the stem question was based on past experience that reinterview respondents often deny stem questions that they endorsed in the baseline interview, leading clinical reinterviewers to declare that the diagnoses based on the initial structured interviews were invalid without clinically reappraising the symptoms reported in the structured interview (Robins 1985; Bromet et al. 1986). We know, however, that the vast majority of respondents who are presented with the fact that they previously endorsed the stem question will confirm this report on reinterview (McLeod et al. 1990), leading to the strong suspicion that failure to obtain high consistency in conventional test–retest reliability studies is due to respondents reluctance to rehash previously reported material more than to memory failure. Some previous reliability and validity 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 reporting inconsistencies (Manuzza et al. 1989; Williams et al. 1992). We rejected this option because of 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 endorsing the stem questions in the earlier interview, this occurred in only a few cases in the NCS clinical reappraisal sample.

Owing to financial constraints we did not include respondents who denied the diagnostic stem questions in the reappraisal study. As a result of this design feature we have no way of estimating the proportion of NCS non-cases who denied the stem questions initially but would have endorsed them and gone on to meet diagnostic criteria on reinterview. It is noteworthy, though, that the test–retest reliability studies in the WHO CIDI Field Trials, reviewed by Wittchen (1994), found that it was rare for respondents who failed to endorse a stem question in an initial CIDI interview to go on to meet full diagnostic criteria for that disorder in a second CIDI interview. Based on this result, the test–retest reliabilities reported below are based on the upper bound assumption that none of the NCS respondents who denied the CIDI stem question in their original interview would have been classified as a SCID case if they had been reinterviewed. This means that the estimated value of the correspondence between the CIDI and the SCID, based on the κ statistic (Cohen 1960; Fleiss 1981), is an upper bound estimate.

Results

Improving question comprehension in the first and second NCS pretests

Most methodological studies of survey instruments start with the problem of question comprehension. Ambiguous questions are obviously more likely to be misconstrued than unambiguous questions. It is less obvious, though, just how large a proportion of questions in most surveys are ambiguous. Belson (1981) was the first survey methodologist to document that many survey questions are misunderstood. Belson carried out a debriefing study of a sample of general population survey respondents in the UK who were given a set of standard survey questions and found that more than 70% of respondents interpreted at least some questions differently from the researcher. Oksenberg, Cannell, and Kalton (1991) came to a similar conclusion in their subsequent debriefing of a nationally representative sample of respondents in the US who were administered standard health interview survey questions. At least one key phrase in two-thirds of the questions in their study was misinterpreted by respondents.

One of the most striking aspects of these studies is that both Belson and Oksenberg et al. found that respondents generally believed that they understood what the investigator meant even when their interpretations of the questions were actually quite idiosyncratic. Oksenberg and her colleagues suggested that this occurs because the survey interview situation is a special kind of interaction in which the standard rules of conversation – rules that help fill in the gaps in meaning that exist in most speech – do not apply. Unlike the situation in normal conversational practice, the respondent in the survey interview often has only a vague notion of the identity of the person to whom he is talking or the purpose of the conversation (Cannell et al. 1968). The person who asks the questions (the interviewer) is not the person who formulated the questions (the researcher), and the questioner is often unable to clarify the respondent's uncertainties about the intent of the questions. Furthermore, the flow of questions in the interview is established prior to the beginning of the conversation, leading to more misreading than in normal conversations, even when questions are seemingly straightforward.

Based on the results of these studies, survey methodologists in recent years have encouraged researchers to carry out cognitive pretest interviews in an effort to learn more about question comprehension. Systematic methods for carrying out these interviews have been developed for this purpose that feature the use of structured debriefing (Ericson and Simon 1993; DeMaio and Rothgeb 1996; Fowler and Cannell 1996). See Sudman et al. (1996), Chapter 2, for a review of these methods. These methods were used in the first two pretests carried out in preparation for the NCS

The entire CIDI was administered in the first of these pretests by experienced interviewers who were trained in the use of the CIDI by the WHO coordinator for CIDI 1.0 (HUW) and a certified CIDI trainer (JMA). Each interviewer completed two practice interviews before beginning the pretest. A random subsample of 20 questions was probed for understanding. As described below, a much smaller proportion of interviewer-rated misinterpretations was found in this pretest than in the Belson and Oksenberg et al. studies. We believe that this favourable result can be attributed to the fact that special care was taken to write CIDI questions in as explicit a fashion as possible. However, as described below, a number of the remaining misinterpretations are systematic and quite serious.

The 30 interviews in the first pretest were tape recorded and the audiotapes coded by a special team of 'behaviour coders' using codes developed by Cannell and his colleagues (Fowler and Cannell 1996) to record three types of behaviour indicative of potential question problems:

- the interviewer misread the question;
- there was a long pause after the question before the respondent answered; and
- the respondent asked for some clarification of the question. Approximately three dozen CIDI questions were pinpointed by this procedure as ones that might be systematically problematic (defined as having at least one problem indicator for at least five respondents).

After the behaviour coding was completed and the results tabulated we debriefed interviewers with a special focus on the potentially problematic questions arising in the debriefing and behaviour coding. The debriefing was carried out using a focus group format in which we went through the entire instrument asking the interviewers to point out questions and procedures that they considered problematic. In addition to eliciting information about the interviewer perceptions of problems, we discussed interviewer impressions about

the reasons for these problems and possible ways they might be resolved.

Respondents were then debriefed in telephone interviews. The focus was on questions considered potentially problematic. Two debriefing methods were used here. The first was the 'think aloud' strategy developed by Ericson and Simon (1993) in which respondents were asked to discuss aloud their thoughts concerning the meaning of questions and how they formulated answers. Respondents were instructed to start 'thinking aloud' as soon as the questions were read to them. The second debriefing method was the 'repeat the question' strategy used by many cognitive psychologists to find out which aspects of questions are most salient to respondents (DeMaio and Rothgeb 1996). This method involves asking respondents to repeat the question they just answered and recording the aspects of the questions that could be recalled and those that could not. This method was used for only 10 questions per respondent based on the finding in previous debriefing studies that respondents become unrealistically attentive to precise wording when they are asked to repeat a larger number of questions.

These debriefing interviews showed that there was considerable confusion about approximately two dozen CIDI questions in addition to those in the schizophrenia section. Three sources of this confusion were inferred. First, a number of the CIDI questions were ambiguously worded. For example, the question about persistence of unrealistic fears of phobic stimuli, which asked respondents whether their fears 'ever lasted months or even years', was commonly misinterpreted as asking about the duration of the autonomic arousal associated with exposure rather than, as was intended, about whether or not autonomic arousal occurred every time the respondent was exposed to the feared object or situation. Changes in question wording were introduced into the revised instrument used in the second pretest to clarify these types of ambiguity. The same two debriefing methods were used in the second pretest as in the first. Much less evidence of miscomprehension was found than in the first pretest. It is noteworthy, however, that the sampling frame was different in the second pretest, which could have accounted for at least some of this difference.

Second, a number of the questions were very complex. For example, the major depression stem question asked about periods of two weeks when 'most of the day most every day' the respondent felt sad or blue or lost interest. It was common to find that respondents who were asked to repeat this question were unable to repeat all important aspects of it (that it had to last a minimum of two weeks, that the depressed mood had to occur nearly every day and that the depressed mood was pervasive on the days it occurred). Some questions were found to create more serious difficulties of this sort than others. For example, none of the respondents who were asked to repeat the following question about agoraphobia was able to recall all important aspects of it:

> When you had this unreasonably strong fear, were you afraid of collapsing, or of the occurrence of other incapacitating or embarrassing symptoms when no help was available or escape possible?

Especially complex questions such as this were broken up into a series of less complex component questions to deal with this problem. In cases where it was important not to break the questions up in this way, we dealt with the complexity problem by instructing interviewers to read the questions slowly. 'Repeat the question' probes in the second pretest revealed that these changes led to a substantial increase in the ability to reproduce all important aspects of questions that were found to be problematic in this respect in the first pretest. It should be noted, though, that confusions still remained in some questions due to complex words and concepts. A decision was made not to modify the questions to address these remaining confusions because this would require us to make more substantial changes to the wording of CIDI questions than we were prepared to make at that time.

Our greatest concern about remaining areas of confusion was with the schizophrenia section. Positive responses to a question about 'hearing things other people couldn't hear', for example, might reflect systematic misunderstanding aimed at normalizing what respondents could perceive as a strange question with a response such as 'yes, I have very good hearing'. The developers of the CIDI were aware of this problem and gave instructions to interviewers to clarify the meaning of the question if the respondent appeared to misunderstand. However, this instruction put a heavy burden on the interviewer to somehow intuit when misunderstanding was occurring. It also created a situation in which active respondents who verbalised their confusion about the meanings of confusing questions obtained more clarification than less active respondents.

We attempted to resolve some of these problems by

developing explicit clarifications that would be provided to all respondents for questions that we judged to be especially confusing, including those in the schizophrenia section. However, concerns were raised by our WHO advisor that these clarifications might alter the meanings of the questions so much that the instrument would no longer be equivalent to the WHO version of the CIDI. These concerns prompted us to remove these clarifications from the version of the instrument used in the NCS. As it happened, our concerns were confirmed in the production interviewing, where we found implausibly high rates of endorsement of questions about hallucinations. Based on this result, we decided to abandon the CIDI as the basis for assessing schizophrenia in favour of a separate clinical reappraisal interview that used the CIDI schizophrenia questions as a first-stage screen (Kendler et al. 1996).

Third, the debriefing also revealed that substantial confusion arose from respondents' failure to understand the purpose of the questions. For example, many pretest respondents misinterpreted the intent of such recall questions as 'In your lifetime, have you ever had two weeks or more when nearly every day you felt sad, blue, depressed?' The misinterpretation concerned the task itself rather than the meanings of words or phrases. About half the respondents in the first pretest interpreted this question as it was intended by the authors of the CIDI, namely, as a request to engage in active memory search and report episodes of the sort in the question. The other half, however, interpreted the question as a request to report whether a memory of such an episode was readily accessible. These latter respondents did not believe that they were being asked to engage in active memory search and did not do so. Not surprisingly, these respondents were much less likely to remember lifetime episodes than those who understood the intent of the question.

The survey methodology literature provides insights into this problem. As noted by Clark and Schrober (1992) in their analysis of discourse rules in survey interviews, the interaction flow in most surveys reinforces the perception that careful response is unimportant. Specifically, normal rules of conversation require a person who is asked a question to signify recognition of turn-taking either by answering the question or by making some other relevant comment (such as, 'um, let me see now . . .') within about one second after the question is issued (Jefferson 1989) unless there is an explicit instruction on the part of the questioner to the contrary. Based on this implicit rule, when an interviewer asks a question that requires considerable thought, the respondent is likely to assume in the absence of instructions to the contrary that the interviewer is operating under normal conversational rules and, as such, is really asking for an immediate and superficial answer.

Cannell et al. (1981) found that this conversational artefact could be minimized by explicitly instructing respondents to answer completely and accurately. Based on this result, we added clarifying statements throughout the modified CIDI developed for the second pretest. The aim here was to inform respondents that accuracy was important. For example, we introduced the CIDI stem questions with the following statement in the second pretest: 'The next question might be difficult to answer because you will need to think back over your entire life. But it is important for our research that you give accurate answers. So please take your time and think carefully before answering.' We found that the use of this introduction resulted in respondents taking a longer time to answer the questions and in a higher proportion of respondents providing positive answers to recall questions. Debriefing showed that this was because respondents were more likely to engage in active memory search rather than estimate in the second pretest than the first.

Motivating accurate reporting in the second NCS pretest

As noted above, the first pretest was concerned largely with question comprehension, whereas the second pretest evaluated a revised version of the CIDI that attempted to correct the comprehension problems discovered in the first pretest. The second pretest also had a second important purpose. This second purpose was to evaluate a series of strategies aimed at addressing another anticipated problem that arose when respondents were clearly told that we needed them to labour at a series of demanding and potentially embarrassing recall tasks: that some respondents would refuse this task either explicitly or implicitly.

Recognition of this potential problem has led survey methodologists to develop motivational techniques to increase the chances that respondents will accept the job of answering completely and accurately. Three techniques that have proven to be particularly useful in this regard in previous methodological studies were used in the second pretest: the use of motivational components in instructions, the use of commitment probes, and the use of contingent reinforcement strategies embedded in interviewer feedback probes.

Motivational instructions

There is evidence that respondents are more willing to undertake laborious and possible painful memory searches if they recognize some altruistic benefit of doing so (Cannell et al. 1981). Even such an uncompelling rationale as 'it is important for our research that you take your time and think carefully before answering' appears to have motivational force. This is even more so when instructions include statements that have universalistic appeal, such as 'accuracy is important because social policy makers will be using these results to make decisions that affect the lives of all of us'. These principles were used in the revisions of the CIDI for the second NCS pretest. We developed an introduction that described the survey as one that would yield important results for health policy makers. We made it clear in question introductions throughout the instrument that we expected thoughtful responses. And we included motivational probes throughout the instrument to let respondents know that 'it is important for our research' that answers be as thoughtful and as accurate as possible. Consistent with our prediction that these changes would improve motivation, we found that the prevalences of reporting lifetime diagnostic stem questions increased in the second pretest compared to the first.

Commitment questions

Previous methodological research has shown that instructions that define the nature of interviewer expectations for respondent behaviour help establish a perspective on the interview that can have motivational force. One way this can be done is by asking an explicit commitment question. Experimental studies carried out by Cannell and his associates (Cannell et al. 1981; Oksenberg et al. 1979a, 1979b) have shown that commitment questions improve accuracy of recall. Furthermore, the work of Cannell and associates (Oksenberg et al. 1979a, 1979b; Cannell et al. 1981) indicates that the joint use of motivating instructions, contingent feedback (described below), and a commitment question has an interactive effect that increases the intensity of memory search and accuracy beyond the effects of any one component separately. This extends not only to the proportion of respondents in different experimental conditions who recall and report past experiences but also to other indicators of commitment such as amount of detail reported and use of personal records and other outside information sources as memory aids during the course of the interview. Based on these results, we used a commitment probe in the second NCS pretest. This was done by prefacing the section of the interview that asked lifetime diagnostic stem questions with the following commitment question:

> This interview asks about your physical and emotional well-being and about areas of your life that could affect your physical and emotional well-being. It is important for us to get accurate information. In order to do this, you will need to think carefully before answering the following questions. Are you willing to do this?

As noted above, we found that the prevalence of diagnostic stem questions increased substantially in the second pretest compared to the first. In addition, consistent with the results of previous studies using similar questions (Cannell et al. 1981), we found that none of the pretest respondents answered the commitment question negatively. Furthermore, when a slightly revised version of this same question was used in the NCS we found that only 35 of the 8133 respondents responded negatively to the commitment question.

Contingent feedback

Several survey researchers have demonstrated that verbal reinforcers such as 'thanks' and 'that's useful' can significantly affect the behaviour of survey respondents. Marquis and Cannell (1969), for example, showed experimentally that the use of such reinforcers resulted in a significant increase in the number of chronic conditions reported in response to an open-ended question about illnesses. These feedback remarks are often used in an unsystematic way, however, as part of general procedures to build and maintain rapport rather than in a systematic way to reinforce good respondent performance.

Based on these observations, Cannell and his associates developed a method for training interviewers to use systematic feedback – both positive and negative – to reinforce respondent effort in reporting (Oksenberg et al. 1979a). The central feature of this method is the use of structured feedback statements coordinated with the content and timing of instructions aimed at reinforcing respondent performance. It is important to recognize that it is performance that is being reinforced rather than the content of particular answers. For example, as noted above, a difficult recall question may be prefaced with the instruction 'This next question may be difficult, so please take your time before answering.'

In contingent feedback, interviewers issue some expression of gratitude whenever the respondent seems to consider his or her answer carefully, whether they remembered anything or not. This structured feedback is programmed periodically throughout the interview in order to maintain the focus on performance standards and to reinforce motivation.

Experiments carried out by Cannell and his associates (Miller and Cannell 1977; Vinokur et al. 1979) have documented that the combined use of these contingent reinforcement probes with instructions explaining the importance of careful and accurate reporting leads to substantial improvement in recall of health-related events in general population surveys, including validated dates of medical events. Importantly, their results also show that self-enhancing response biases are reduced when these strategies are used, as indicated by both a decreased tendency to under-report potentially embarrassing conditions and behaviours (such as gynecological problems, seeing an X-rated movie) and a decreased tendency to over-report self-enhancing behaviours (such as the number of books read in the last three months or reading the editorial page of the newspaper the previous day). Based on these results, interviewers in the second NCS pretest were trained and instructed to use contingent feedback to reinforce thoughtful reporting.

Facilitating accurate reporting

In addition to considering strategies that improve question comprehension and motivation to report accurately, the literature on survey methodology has explored the use of methods to improve accuracy once respondents commit to active memory search. The latter techniques provide recall aids that increase the efficiency of memory work. Our main focus was on improving accuracy of recall in answering lifetime diagnostic stem questions. A good deal of evidence suggests that people who answer negatively to 'Have you ever . . .?' questions (Shannon 1979; Glucksberg and McCloskey 1981) usually do so based on a lack-of-knowledge inference (Genter and Collins 1981); that is, they answer negatively based on a conclusion drawn from lack of immediate recollection of an experience.

There are a number of processes that can bring about this inference when it is not true. One of these considered in the NCS involved the pace of the interview. A number of survey methodologists have noted that unless interviewers are carefully trained to the contrary, they will ask questions too quickly, which will reduce the accuracy of respondent reports (Neter and Waksberg 1964; Cannell et al. 1977; Sudman and Bradburn 1982). This is especially true for lifetime recall questions. At least two fairly obvious processes are involved here. Haste on the part of the interviewer conveys the message that a quick response is more important than an accurate response (Clark and Schober 1992). Memories are also more likely to be recovered when respondents are allowed to think at their own pace rather than being rushed (Bradburn et al. 1987).

These observations and the analysis of interaction sequences in interviews have led Cannell and his associates to recommend that the reading pace of the interviewer should be no more than an average of two words per second (Cannell et al. 1981), that respondents should be explicitly asked to think at their own pace (Cannell and Kahn 1968) and that critical questions should be designed to encourage periods of silence that are explicitly defined as thinking time (Cannell 1985a). Several experiments have documented that these procedures lead to more accurate recall of health-related events (Lessler et al. 1989; Burton and Blair 1991; Means et al. 1993).

The second NCS pretest used the results of these previous methodological studies to have interviewers read all diagnostic stem questions slowly and deliberately and to follow these questions with the request 'Please take your time and think carefully before answering.' The proportion of respondents who endorsed diagnostic stem questions increased dramatically compared to the first pretest. Based on this encouraging result, the third pretest expanded the strategy by developing a 'lifetime review section' shortly after the beginning of the interview. This section began with the commitment probe described earlier in this paper followed by a general injunction for respondents to take their time and think carefully. This was then followed by the slow and deliberate reading of the diagnostic stem questions for all the sections of the CIDI.

Our reasoning in developing the lifetime review section was that this special section, administered close to the beginning of the interview, would catch respondents when they were mentally fresh and motivated to carry out an active memory search for lifetime episodes of the disorders. In addition, by asking all the diagnostic stem questions before respondents became

aware that each affirmative response would result in them being asked a great many additional questions, we hoped to avoid the negative response set that has been shown to occur when respondents who want to reduce interview length become aware of stem-branch sequences (Bradburn et al. 1979). As described below, a subsequent experiment verified that this approach leads to a dramatic increase in the proportion of respondents endorsing diagnostic stem questions.

Acknowledging the limits of reporting ability

Research on basic cognitive processes has shown that memories are organized and stored in structured sets of information packages commonly called *schemas* (Markus and Zajonc 1985). When the respondent has a history of many instances of the same experience that cannot be discriminated, the separate instances tend to blend together in memory to form a special kind of memory schema called a *semantic memory*, a general memory for a prototypical experience (Jobe et al. 1990; Schwarz 1990; Means and Loftus 1991). For example, a person may have a semantic memory of what panic attacks are like but, due to the fact that he has had many such attacks in his or her lifetime, may not be able to specify details of any particular panic attack. In comparison, when the respondent has had only a small number of lifetime experiences of a certain sort or when one instance stands out in memory as much different from the others, a memory can probably be recovered for that particular episode. This is called an 'episodic memory'.

The effects of memory schemas and the difference between semantic and episodic memories are central themes in research on autobiographical memory. Indeed, we must determine whether episodic memories can be recovered and whether the respondent is answering the questions by referring to episodic memories or by drawing inferences of what the past must have been like on the basis of more general semantic memories. Research shows that people are more likely to recover episodic memories for experiences that are recent, distinctive, and unique, while for experiences that are frequent, typical, and regular, people will rely more on semantic memories (Brewer 1986; Belli 1988; Menon 1994).

When a survey question is designed to ask about a particular instance of an experience, it helps accuracy of data collection if the question is posed in such a way that the respondent knows he or she is being asked to recover an episodic memory. In order to do this, though, the researcher needs to have some basis for assuming that an episodic memory can be recovered for the experience. If it cannot, a question that asks for such a memory implicitly invites the respondent to infer or estimate rather than remember and this can have adverse effects on quality of reporting later in the interview (Pearson et al. 1992). In comparison, it should be made clear whether a question is designed to recover a semantic memory or to use semantic memories to arrive at an answer by estimation.

One difficulty with these injunctions in the case of retrospective recall questions about lifetime psychiatric disorder is the uncertain level of recall accuracy. We confronted this problem in the first NCS pretests when we asked the standard CIDI questions about first onset: 'When was the first time you had [disorder]?' The debriefing of pretest respondents revealed that whereas some people had very vivid memories of their first onsets, others had no such memory. The problem posed by this variation was how to develop a method of asking the question that reinforced our overall commitment to collecting complete and accurate information, while simultaneously recognizing the limits of autobiographical memory and avoiding a request for a precise answer from the subsample of respondents who were unable to recover an episodic memory for their first episode.

We resolved this problem by adapting several of the principles discussed above in a three-part question series designed to inform respondents that answers should be as precise as possible while still recognising the limits of memory. The question sequence began with what has been referred to in the methodology literature as a 'prequest' – a question aimed at clarifying the nature of the request for information in subsequent questions. The prequest question was:

> Can you remember your EXACT age the VERY FIRST TIME you had a sudden spell of feeling frightened or anxious and had several of these other things ['other things' refers to a checklist of symptoms that respondents previously reported which was presented for visual review on a cue card] at the same time? [Emphasis in original.]

During the second pretest we probed positive responses to determine the basis for exact recall and discovered that, overall, these respondents were either younger (so the event was likely to have occurred more recently), had a smaller number of lifetime episodes, or had a distinctive context that allowed them to date the

age of their vividly recalled first attack. Based on this information, the final question series simply followed this answer with the question 'How old were you?' In comparison, respondents who answered the prequest negatively were asked a different follow-up question phrased in such a way as to make it clear that we wanted an estimate, based on our understanding that the respondent could not provide an exact answer. This question was: 'ABOUT how old were you [the first time you had one of these attacks] [emphasis in original]?' Interviewers were instructed to accept a range response (such as 'sometime in my early 20s') without probing, as we were soliciting an estimate. This question was then followed by another that was designed to provide an upper bound on our uncertainty concerning age of onset and to permit the respondent to answer even when uncertain about the exact age of the first attack: 'What is the earliest age you can CLEARLY REMEMBER having one of these attacks?' (Emphasis in original.)

The latter question is much less demanding than the original question about the exact age of the very first attack, and, not surprisingly in light of this, virtually all respondents in the second pretest and in the subsequent NCS were able to provide an age in their answer. Interestingly, the age given in response to this question was often younger than the lower bound of the age range given in response to the preceding question. This result was noticed in analysing the data from the second pretest. Therefore, in the third pretest, respondents with this pattern were asked to explain the discrepancy. Their responses suggested that this seemingly inconsistent result was due to the fact that estimation was typically used to arrive at the response to the question 'ABOUT how old were you . . .?' while active memory search focusing on the part of the lifespan implied by the answer to the preceding question was used to arrive at the response to the subsequent question.

An experimental comparison of the WHO CIDI and the modified CIDI

As noted above, comparison of the responses in the first and second pretests suggested a reduction in miscomprehension and an increased effort to answer accurately as a result of the CIDI changes. However, as there was no experimental manipulation of these conditions and several aspects of the two pretests differed in ways that are potentially important for response quality (such as method of recruitment, site of interview administration, focus of probing), it was impossible to draw firm conclusions about the effects of these changes from a comparison of the two pretests. As a result, a split ballot experiment was carried out in which respondents were randomly assigned to receive either the original WHO CIDI 1.0 or the modified version of the CIDI developed in the three pretests. This experiment focused on the impact of these changes on stem question and diagnosis endorsement probabilities.

Summary results are reported in Table 1 for each of the seven CIDI mood and anxiety disorders assessed in the experiment. The first entry for each disorder is the proportion of respondents in each condition who endorsed the diagnostic stem question. The modified procedures resulted in an increased proportion of respondents endorsing diagnostic stem questions for six of the seven disorders. Three of these six differences are significant at the 0.05 level, each of them involving an extreme version of normal mood variation that one might expect to require serious memory search to recover (two weeks of depressed mood or anhedonia, six months of worry, a fear of leaving home alone or of being alone away from home). The modified version of the instrument did not have any statistically significant effects, in comparison, on the proportion of respondents reporting the more vivid experiences associated with panic, extreme fear of phobic stimuli, or long periods of depressed mood lasting two years or longer.

The second entry for each disorder in the table shows the proportions of respondents in each condition who met full lifetime criteria among those who endorsed the stem question for the disorder. The WHO CIDI diagnostic computer program was used to generate diagnosis prevalence estimates. The modified procedures resulted in increased conditional proportions of respondents who met criteria for six of the seven disorders. Two of these six differences, those for agoraphobia and for simple phobia, are statistically significant at the 0.05 level.

The last two rows in the table report the proportions of respondents who met lifetime criteria for at least one of the seven disorders and for more than one of these disorders. Nearly twice as many people who were given the modified version of the instrument (38.7%) met CIDI criteria for at least one disorder, as compared with the original version (21.0%). Even more strikingly, nearly four times as many respondents who were given the modified (19.0%) as the original (4.8%) CIDI met criteria for two or more disorders.

Table 1: Estimated lifetime prevalences of diagnostic stems and full diagnostic criteria in an experimental comparison of the WHO CIDI and modified CIDI.

DSM – III – R Diagnoses		WHO CIDI %	(se)	Modified CIDI %	(se)	t-test
I. Mood disorders						
Major depressive episode	Stem	28.6	(4.4)	51.9	(4.9)	3.5[2]
	disorder/stem	33.3	(8.9)	47.3	(7.6)	1.2
Dysthymia	Stem	14.3	(3.4)	17.0	(3.6)	0.5
	disorder/stem	40.0	(12.6)	50.0	(11.8)	0.6
Any mood disorder[1]	Stem	32.4	(4.6)	52.8	(4.8)	3.1[2]
	disorder/stem	35.3	(8.2)	50.0	(6.7)	1.4
II. Anxietydisorders						
Panic disorders	Stem	17.1	(3.7)	26.4	(4.3)	1.6
	disorder/stem	5.8	(5.5)	7.2	(4.9	0.2
General anxiety disorder	Stem	3.8	(1.9)	20.8	(4.0)	3.8[2]
	disorder/stem	25.0	(21.6)	31.8	(9.9)	0.3
Agoraphobia	Stem	4.8	(2.1)	19.8	(3.9)	3.4[2]
	disorder/stem	20.0	(17.9)	76.2	(9.3)	2.8[2]
Simple phobia	Stem	42.9	(4.8)	53.8	(4.8)	1.6
	disorder/stem	6.7	(3.7)	19.3	(5.2)	2.0[2]
Social phobia	Stem	32.4	(4.6)	28.3	(4.4)	0.6
	disorder/stem	29.4	(7.8)	26.7	(8.1)	0.2
Any phobia	Stem	56.2	(4.8)	64.2	(4.7)	1.2
	disorder/stem	22.0	(5.4)	39.7	(5.9)	2.2[2]
Any other anxiety disorder	Stem	19.0	(3.8)	46.2	(4.8)	4.4[2]
	disorder stem	10.0	(6.7)	16.3	(5.3)	0.7
III. Mood or anxiety disorders						
Any disorder		21.0	(3.9)	38.7	(4.7)	2.9[2]
Two or more disorders		4.8	(2.1)	19.0	(3.8)	3.3[2]
(n)		(105)		(106)		

[1]Diagnoses were made without using DSM-III-R hierarchy rules.

[2]The prevalence estimates in the two subsamples differ significantly at the 0.05 level.

Apparently, a sizeable number of respondents avoided providing an answer that was likely to elicit numerous follow-up questions once they became aware of this contingency. It is therefore important to administer all stem questions before this contingency becomes apparent. This last result suggests that the use of the life review section at the beginning of the interview had its greatest effect on responses to subsequent stem and symptom questions among respondents who had already been administered the full set of questions for one or more diagnostic sections.

Comparison of the NCS and ECA prevalence estimates

As one would expect from the results of the experiment, the NCS produced much higher lifetime prevalence estimates for most disorders than those found a decade earlier in the ECA Study (Robins and Regier 1991). Overall, 48% of NCS respondents were estimated to meet lifetime criteria for at least one DSM disorder (Kessler et al. 1994) compared with 32.7% of ECA respondents (Robins et al. 1991). The ECA diagnoses were generated from the Diagnostic Interview

Schedule (DIS) (Robins et al. 1981), a fully structured interview on which the subsequent development of the CIDI was based and that shared the CIDI comprehension and motivation problems discovered in the NCS pretests.

As noted previously in the section on methods and procedures, a subsequent *post hoc* comparison of the ECA and NCS was carried out in collaboration with Darrel Regier and his staff at NIMH. The ECA-NCS comparison was designed to determine how much of the dramatic difference between the prevalence estimates in the two surveys might be due to differences in sampling (local samples in each of five largely urban areas in the ECA compared to a nationally representative sample in the NCS), age range (18+ in the ECA compared with 15–54 in the NCS), differences in depth of questions, or diagnostic system (DSM-III in the ECA compared to DSM-III-R in the NCS).

As described more fully by Regier et al. (1997), we compared the subsample of 18–54 year old non-Hispanic whites in the ECA with those in the urbanized NCS subsample and limited the analysis to the subset of diagnoses that were assessed in a roughly comparable way in the two surveys. This second restriction allowed us to control for the fact that the CIDI includes more items to assess certain diagnoses than the instrument used in the ECA study. The age range was constrained in order to correct for the exclusion of elderly people in the NCS. The geography was constrained in order to deal with the undersampling of rural areas in the ECA. The focus on non-Hispanic whites was made in order to adjust for the fact that the minorities in the ECA were unrepresentative of those in the total US (the Hispanics almost exclusively being represented by Mexican-Americans residing in one neighbourhood in Los Angeles and the blacks over-representing those from low income urban areas). Finally, all NCS diagnoses were redefined using DSM-III criteria.

Not suprisingly in light of the experimental results reported above, these structural modifications did not totally explain the discrepancies between ECA and NCS lifetime prevalence estimates (Regier et al. 1997), although they did meaningfully reduce them. The question arises of whether the remaining discrepancies are due to some combination of four processes: residual differences in NCS and ECA assessment procedures that could not be controlled in our analysis, overestimation of lifetime prevalences in the NCS, underestimation of lifetime prevalences in the ECA, and the true lifetime prevalences of DSM-III disorders increasing over the decade between the time the two surveys were fielded. We acknowledge the first of these possibilities, but have no way to correct for it. We dismissed the last of these possibilities based on the fact that retrospective reports by NCS respondents in the same cohorts as ECA respondents yielded higher lifetime prevalence estimates as of the years the ECA was carried out than those obtained in the ECA from respondents in these same cohorts. The remaining possibilities – overestimation in the NCS, underestimation in the ECA, or both – could be major determinants of the prevalence differences between the two surveys.

Further analysis led us to conclude that underestimation in the ECA is a more important factor than overestimation in the NCS. The most direct and persuasive evidence for this conclusion is presented in the next section of the paper, where we review the results of the NCS clinical reappraisal study. As is described in more detail in that section, blind clinical reinterviews documented that the CIDI interviews in the NCS did not overdiagnose lifetime DSM-III-R disorders. However, indirect evidence consistent with this finding can be seen in the ECA-NCS comparisons. Perhaps the most dramatic of these comes from an analysis of the combined lifetime prevalence data estimated in the baseline ECA and the ECA follow-up survey completed one year after baseline. This follow-up survey repeated the lifetime assessment rather than asking only about disorders that occurred during the year between waves. The lifetime prevalence estimates increased dramatically when the two waves were combined. For example, the lifetime prevalence estimate of any DSM-III disorder in the combined sample was 46.9%, very similar to the 48% obtained in the NCS (Regier et al. 1997). For some individual disorders the prevalence increases were even more dramatic. For example, there was a 78% increase in the estimated lifetime prevalence of major depression when the two waves of data were combined.

Importantly, the vast majority of the people who reported a disorder in the follow-up of the ECA but not the baseline interview stated in the reinterview that their age of onset was well before the baseline interview. This means that the increase in lifetime prevalence at follow-up is due largely to remembering past episodes rather than to reporting new onsets in

the year between the two waves of data collection. The most plausible interpretation of this finding that the two-wave ECA prevalence estimates approximate the one-wave NCS estimates is that the motivation and memory-enhancing procedures used in the NCS stimulated more complete reporting of lifetime disorders in the NCS than the ECA. This interpretation is consistent with the finding of Jay Turner and his colleagues in Toronto (personal communication, 1997), who carried out a two-wave panel survey with a 12-month time interval between waves using the same interview procedures as in the NCS, that the increase in lifetime prevalence estimates was much smaller than in the ECA Study.

This interpretation is also consistent with the finding that the lifetime prevalence differences between the ECA and NCS are due largely to differences in the probability of endorsing a lifetime diagnostic stem question rather than in the conditional probability of meeting full diagnostic criteria after endorsing the stem question. A good example is the DSM-III diagnosis of major depressive episode, a syndrome having one of the largest discrepancies in lifetime prevalence estimates in the matched NCS (15.5%) and ECA (7.7%) comparison samples. This discrepancy is due to nearly twice as high a proportion of respondents endorsing the stem questions for this syndrome in the NCS (56.8%) as the ECA (32.2%), whereas very similar proportions of respondents met full diagnostic criteria in the subsample of respondents who endorsed the stem questions (27.3% in the NCS and 24.0% in the ECA).

The question could be raised as to whether the higher stem endorsement probability in the NCS than ECA resulted in the NCS focusing on more people with symptom profiles that were not clinically signifi-

Table 2: Distributions of Major Depression Criterion A symptoms in the 'matched' ECA (one wave) and NCS samples[1]

	Total		Black		White		Hispanic	
	ECA %	NCS %	ECA %	NCS %	ECA %	NCS %	ECA %	NCS %
I. Criteria stem endorsement								
A1/2 Depressed mood and/or anhedonia	32.2	56.8	26.0	50.6	33.7	58.3	27.6	53.7
II. Symptom endorsement among respondents who endorsed the stem (either A1 or A2)								
A3. Significant weight/appetite change	56.7	56.4	60.1	54.8	57.0	56.1	48.9	60.2
A4. Insomnia or hypersomnia	55.9	65.0	51.3	64.1	56.6	64.7	53.4	69.2
A5. Psychomotor agitation or retardation	49.9	57.3	42.1	56.5	51.7	57.0	41.1	60.7
A6. Fatigue/loss of energy	31.2	37.3	35.9	38.0	31.3	35.9	23.3	47.5
A7. Feeling of worthlessness or guilt	33.9	39.2	26.9	33.8	35.1	39.4	30.4	43.3
A8. Diminished ability to concentrate	45.4	50.5	41.7	48.6	46.6	50.7	36.4	50.8
A9. Recurrent thoughts of death or suicidality	57.7	62.2	51.2	57.1	58.9	63.0	53.2	61.4
III. Five or more criteria endorsed (including A1 or A2)	24.0	27.3	20.4	26.4	24.9	27.4	20.5	28.6
IV. Rank order correlation of symptom prevalences	0.60		0.81		0.88		0.88	

[1]Urban respondents in the age range 18–54.

[2]The percentages reported in parts II and III of the table are conditional percentages in the subsample of respondents who reported either depressed mood (DSM-III-R) or anhedonia (Criterion A2). For example, the 56.7% of the 32.2% of the Criterion A1/A2 positives who endorsed the CIDI questions about significant weight/appetite change represent 18.3% (0.567×0.322) of the total sample.

cant. This does not appear to be the case from our comparison of NCS and ECA symptom profiles in the subsamples of respondents who endorsed the diagnostic stem questions. An example is presented in Table 2, where we show symptom profiles for major depressive episodes in age-matched NCS and ECA subsamples broken down by race/ethnicity. As shown there, symptom prevalence and distributions are very similar across the surveys in the subsample of those who endorsed the depression stem questions. This suggests that the same disorders were being recalled more completely in the NCS than the ECA rather than that the NCS was eliciting information about less severe syndromes than the ECA.

Clinical reappraisal of the NCS CIDI cases

The ultimate test of the CIDI modifications developed for the NCS would be a validity study. In the ideal case, such a study would assign randomly selected community samples to be interviewed with either the original CIDI or the modified CIDI and then blindly reinterviewed with a gold standard clinician-administered diagnostic interview. A comparative validity study of this sort was beyond the resources of the NCS pretest study budget and was not approved for funding in several subsequent attempts to seek separate support for such a study. However, as noted above, we were able to carry out a modest clinical reappraisal study of lifetime prevalence estimates among NCS respondents who endorsed diagnostic stem questions.

Results are presented in Table 3, where we show the relationships of both the CIDI diagnoses in the NCS and the retest CIDI diagnoses with clinical diagnoses based on the SCID. The first column of the table shows the positive predictive value (PPV) of the diagnoses. This is the proportion of CIDI cases confirmed as cases in the clinical reinterviews. The PPVs for all disorders other than panic disorder and generalized anxiety disorder are greater than 0.50, ranging from 0.60 to 0.74 for mood disorders, from 0.63 to 0.93 for anxiety disorders, and from 0.87 to 1.0 for substance disorders. There is a tendency for values of PPV to be somewhat higher for the retest CIDI than the CIDI diagnoses in the NCS, indicating that part of the discrepancy between the NCS CIDI and the SCID is due to inconsistency in respondent reports over time.

Clinical diagnoses of CIDI non-cases are reported in the negative predictive value (NPV) columns in Table 3. Net predictive value is the proportion of CIDI non-cases confirmed as non-cases in the clinical reinterviews. The proportions of CIDI non-cases with positive stems who were determined to be cases in the clinical reinterviews (1-NPV1) range from 0.00 to 0.20 for mood disorders, from 0.07 to 0.50 for anxiety disorders, and from 0.15 to 0.33 for substance disorders. There is a tendency for values of 1-NPV1 to be somewhat lower for the retest CIDI than the CIDI diagnoses in the NCS, indicating that part of the discrepancy between the NCS CIDI and the SCID is due to inconsistency in respondent reports over time.

We would expect PPV to be considerably larger than 1-NPV1 if the CIDI-discriminated true cases from non-cases in the subsample of respondents who endorsed diagnostic stem questions. The table shows that this is the case. Ratios of the proportions of CIDI cases versus non-cases that were classified as cases in the SCID are greater than 5:1 for seven of the 19 entries in the table, between 2:1 and 5:1 for 10 others, and 2:1 or less for the remaining two entries. Positive predictive value is significantly greater than 1-NPV1 at the 0.05 level in 18 of the 19 comparisons in the table, the one exception being for the retest of generalised anxiety disorder. The next entries in the table, NPV2, are estimated upper bounds of NPV based on the assumption that none of the CIDI stem-negative non-cases would have been classified as cases if they had received clinical reappraisal interviews. The values of these entries range from 0.91 to 1.0 for mood disorders, from 0.85 to 1.0 for anxiety disorders, and from 0.67 to 0.85 for substance disorders.

The κ values linking the CIDI and SCID diagnoses are reported in the next column of the table; κ was computed on weighted data. The computations adjusted for the oversampling of CIDI cases in selecting respondents for clinical reappraisal interviews and, as noted above, were based on the assumption that none of these respondents would have been classified as cases if clinical reinterviews had been carried out among them. The κ values for the NCS CIDI are above 0.6 for two diagnoses (agoraphobia, social phobia), between 0.5 and 0.6 for five others (major depressive episode, mania, simple phobia, alcohol-use disorders, and drug use disorders), and below 0.5 for two others (0.43 for panic disorder and 0.35 for generalised anxiety disorder). There is a general tendency for the κ values to be somewhat higher for the retest CIDI than the CIDI diagnoses in the NCS, with five of the κs above 0.60 and two others close to 0.60.

Table 3: Concordance between the CIDI and SCID in diagnosis-specific subsamples of NCS respondents

DSM-III-diagnoses[1]	Time of CID[6]	Positive predictive value			Negative predictive value						K		Bias[9]	
		PPV	(se)	(n)	1-NPV1	(se)	t-test[7]	NPV2	(se)	(n)	K	(se)	+/–	χ^2
I. Mood disorders														
Major depressive episode	N	0.65	(0.11)	(20)	0.20	(0.13)	2.7•	0.91	(0.09)	(10)	0.53	(0.14)	–	0.0
	R	0.60	(0.10)	(25)	(0.00)	(–)	3.7[8]	1.0	(–)	(5)	0.71	(0.11)	+	8.0**
Mania[2]	N	0.74	(0.16)	(7)	0.08	(0.09)	3.6•	0.99	(0.01)	(52)	0.58	(0.11)	–	1.8
II. Anxiety disorders														
Panic disorder	N	0.47	(0.11)	(19)	0.10	(0.09)	2.6•	0.98	(0.04)	(10)	0.43	(0.38)	–	0.4
	R	0.54	(0.14)	(13)	0.19	(0.10)	2.0•	0.95	(0.05)	(16)	0.36	(0.34)	–	2.4
General anxiety disorders	N	0.21	(0.11)	(20)	0.30	(0.13)	1.6	0.97	(0.02)	(10)	0.35	(0.26)	–	1.7
	R	0.23	(0.13)	(29)	0.21	(0.10)	2.0•	0.98	(0.01)	(14)	0.47	(0.24)	+	0.2
Agoraphobia	N	0.64	(0.20)	(22)	0.17	(0.15)	1.9•	0.98	(0.01)	(6)	0.63	(0.18)	+	0.0
	R	0.92	(0.11)	(13)	0.20	(0.10)	4.8•	0.98	(0.02)	(15)	0.79	(0.13)	–	0.0
Simple phobia	N	0.83	(0.09)	(24)	0.36	(0.15)	2.7•	0.85	(0.04)	(11)	0.54	(0.11)	–	3.2
	R	0.79	(0.10)	(28)	0.29	(0.17)	2.5•	0.88	(0.04)	(7)	0.57	(0.11)	–	4.2**
Social phobia	N	0.95	(0.05)	(22)	0.50	(0.16)	3.0•	0.88	(0.04)	(9)	0.68	(0.09)	–	1.5
	R	0.93	(0.06)	(27)	0.00	(–)		1.0	(–)	(4)	0.95	(0.04)	–	0.1
Post traumatic stress disorder[3]	N	0.67	(0.17)	(19)	0.11	(0.03)	3.2•	0.89	(0.03)	(11)	0.39	(0.16)	–	1.5
	R	0.93	(0.06)	(14)	0.07	(0.05)	11.0•	0.93	(0.02)	(16)	0.66	(0.15)	–	0.1
III. Substance disorders														
Alcohol abuse/dependence[4]	N	0.92	(0.06)	(20)	0.25	(0.14)	4.5•	0.75	(0.14)	(10)	0.54	(0.09)	–	1.3
	R	1.0	(–)	(19)	0.15	(0.11)	7.2[8]	0.85	(0.11)	(11)	0.74	(0.07)	–	0.2
Drug abuse/dependence[5]	N	0.87	(0.07)	(25)	0.33	(0.19)	2.7•	0.67	(0.14)	(6)	0.39	(0.12)	–	2.4
	R	1.0	(–)	(29)	0.14	(0.14)	3.5[8]	0.86	(0.11)	(2)	0.59	(0.11)	–	0.1

*PPV is significantly larger than 1-NPV1 at the 0.5 level.

**The estimated prevalence based on the CIDI is significantly different from the estimated prevalence based on the SCID at the 0.5 level.

[1]Diagnoses were made without using DSM-III-R hierarchy rules.

[2]No retest CIDI data are reported because omission of one criterion in the reinterview made it impossible to generate a CIDI retest diagnosis of mania.

[3]All computations are based on the subsamples of respondents who reported trauma exposure.

[4]All computations are based on the subsample of respondents who reported ever drinking one or more drinks in a single year of their life.

[5]All computations are based on the subsample of respondents who reported ever using drugs one or more times in their life.

[6]Results in rows marked 'N' compare the CIDI assessed in the NCS with the SCID, while results in rows marked 'R' compare the CIDI assessed in the reappraisal interview with the SCID.

[7]t-tests evaluate the significance of differences between PPV and 1 – NPV1.

[8]Significance tests of difference scores were computed by adding 0.5 to the subsamples with no variance in the prevalence estimates to generate standard errors of these estimates.

[9]Chi-square tests evaluate the significance of differences between estimated prevalences based on the CIDI and the SCID using the McNemar procedure.

Values of less than 1.0 for κ mean that there is imprecision in the CIDI prevalence estimates. It is not clear, though, whether there is bias, as both false positives (PPV less than 1.0) and false negatives (NPV2 less than 1.0) were observed. Bias was evaluated by comparing the CIDI estimated prevalences with the SCID estimated prevalences based on two-by-two tables of weighted data that took into consideration the unequal probabilities of selection into the clinical reinterview samples. The McNemar test (Bishop et al. 1975) was used to test the significance of prevalence differences. As shown in the last column of Table 3, the majority of the CIDI prevalence estimates were lower than the SCID prevalence estimates (16 of 19). However, none of the differences involving the CIDI prevalence estimates in the NCS is statistically significant at the 0.05 level. Two involving the retest CIDI are significant, one of them an overestimated prevalence of major depressive episodes and the other an underestimated prevalence of simple phobia in the CIDI compared to the SCID.

Conscious non-disclosure

It must be noted that we ignored the difficult question of whether respondents are honest either with the interviewer or with themselves in discussing their history of psychopathology. The issue of honesty is a problematic one. The methodological literature on the accuracy of respondent reports shows clearly that the perceived social desirability of responses is important in determining the accuracy of reports (Sudman and Bradburn 1974; Kessler and Wethington 1991), although recent research suggests that the magnitude of social desirability bias is smaller than previously suspected (Sudman et al. 1996). We have no way of assessing the magnitude of this problem in the CIDI with available data, but it clearly needs to be taken into consideration as methodological refinements of the CIDI continue to be made. Strategies exist to increase willingness to disclose embarrassing information (Bradburn et al. 1979; Turner et al. 1992). The use of these strategies should be investigated in future methodological studies of the CIDI.

Future directions

Improving question comprehension

The NCS pretest study results clearly showed that our efforts to improve question comprehension were useful. A number of the wording changes implemented in the NCS on the basis of the pretests were subsequently introduced into the most recent version of the CIDI, Version 2.1 in order to improve comprehension. For example, the very complex question in CIDI 1.0 concerning agoraphobic fears – 'Were you afraid of collapsing, or of the occurrence of other incapacitating or embarrassing symptoms when no help was available or escape possible?' – was changed in CIDI 2.1 to a pair of much less complex questions:

> Were you afraid of [SITUATIONS DESCRIBED EARLIER] because you would be unable to escape if you suddenly had some of these problems?
>
> Were you afraid of [SITUATIONS DESCRIBED EARLIER] because you might be unable to get help if you suddenly had some of these problems?

Our sense is that the most recent English language version of the CIDI has few remaining basic comprehension problems, although further cognitive interviewing with representative community samples will be needed to verify this impression. The remaining potential problems of this sort that we have been able to extract from our review of the instrument mostly involve three sorts of questions. The first are questions that rely on the understanding of a single focal word, such as 'During that period, were you restless?' This question is difficult because it requires the respondent to understand the word 'restless' and also to infer something about the level and persistence of the symptom that would qualify as being enough to mention. There are some questions in CIDI 2.1 that address both of these problems in secondary clauses. For example, the following question in the somatization section about paralysis offers both a definition of the term and a clear specification of minimal level and persistence: 'Have you ever been paralysed – that is, completely unable to move a part of your body for at least a few minutes?' The decision to include clarifications such as this in some questions but not others was based entirely on the intuition of the person writing the question regarding whether most people would understand the question without such clarifications. We believe that the instrument would be improved by basing decisions of this sort on empirical evidence against a gold standard (the determination whether the question wording yields responses that are consistent with clinician symptom evaluations) rather than intuition.

The second set of potentially problematic questions remaining in the CIDI from the point of view of comprehension are those that rely on vague quantifiers. An

especially important example that appears repeatedly throughout the CIDI is the question 'Did [SYMPTOM] interfere with your life and activities a lot?' Methodological studies show that there is enormous between-person variability in the interpretation of vague quantifiers such as the word 'a lot' (Schaeffer 1991). Our own preliminary studies, in fact, suggest that only about half as many respondents report 'a lot' of interference due to psychiatric problems in response to the question 'How much did (SYMPTOM) interfere with your life or activities – a lot, some, a little, or not at all?' than if they are asked a yes–no question about interfering a lot. Consistent responses to such questions are critical to the validity of the CIDI. As a result, further methodological research is needed to determine whether the use of vague quantifiers in questions of this sort can be justified and, if not, whether more explicit questions can be developed to avoid the use of vague quantifiers.

The third set of potentially problematic questions remaining in the CIDI from the point of view of comprehension are those that use explicit quantifiers embedded in questions that are so complex that the quantifiers are not heard by all respondents. A good example is the CIDI stem question for major depression:

> Now I want to ask you about periods of feeling sad, empty, or depressed. In your lifetime, have you ever had two weeks or longer when nearly every day you felt sad, empty, or depressed for most of the day?

As noted earlier in the paper, the NCS pretests found that many respondents do not hear all aspects of this question. Indeed, in more recent methodological studies, we have found that as many as one-fourth of the respondents who endorse this stem question contradict themselves in response to one or both of the following two additional probes:

> On the days you felt this way, did these feelings last all day long, most of the day, about half the day, or less than half the day?
>
> During these two weeks, did you feel this way every day, almost every day, or less often than that?

Such discrepancies demonstrate that many respondents attend to some aspects of complex questions more than others. In cases when it is important that all aspects of the question be heard, as in this example, it might be necessary to ask a series of separate focused questions rather than a single question. Future methodological research is needed to investigate this possibility.

Improving question calibration in relation to diagnostic criteria

An issue related to the vague quantifier problem concerns the possibility that some CIDI questions are not calibrated to the same level of intensity as the DSM or ICD criteria they are designed to operationalize. This is not a problem of respondent comprehension but of mismatch between the naive psychometrics implicit in the CIDI question-writing process and the validation standard the questions are intended to approximate. For example, the DSM-IV Criterion A3 for substance abuse of 'recurrent substance-related legal problems (e.g. arrests for substance-related disorderly conduct)' is operationalized in the CIDI with the question 'In your lifetime, has your use of [substance] ever led to problems with the police?' Aside from the fact that this question does not ask about 'recurrent' problems, there is the possibility that the word 'problem' means something quite different to different people depending on their history of contact with the legal system. Variation of this sort, if it exists, could presumably be corrected without great difficulty by clarifying the types of experiences that qualify as problems.

Investigations of this sort will require a comparison of CIDI responses to some validation standard. We have carried out preliminary work of this sort in the NCS clinical reappraisal sample in which we compared criterion-level responses to the CIDI with the SCID for several diagnoses (Wittchen et al. 1995, 1996). Our overall finding was that criterion-level agreement was generally quite good and that serious calibration problems were limited to a small number of criteria. In the case of generalized anxiety disorder, for example, we found that there was only one major threshold problem: a disagreement between the CIDI and the clinical interviewers on criterion A2 involving whether the worries were excessive or unrealistic.

Replicated criterion-level investigations of this sort across a wide range of populations are needed to pinpoint and correct systematic calibration problems in the CIDI. We are currently involved in a cross-national collaborative study of this sort involving the CIDI assessment of post-traumatic stress disorder (PTSD). In this work, we are expanding the structured CIDI questions for all the criteria of PTSD where calibration

problems might exist and using systematic cognitive interviewing to make sure there are no comprehension problems in these new questions. Then we are administering this expanded section of the CIDI to samples of trauma victims around the world by collaborators who are using the same clinical reappraisal interview. Our hope is that this exercise will lead to the documentation of cross-nationally consistent but limited CIDI calibration problems that can be corrected by modifying the assessment of the problematic criteria to use some subset of the expanded CIDI questions included in the data collection. If so, this evaluation of the CIDI PTSD section could serve as a model for future CIDI developments.

Motivating accurate reporting

The evidence reviewed above from the methodological literature is quite clear that there are endemic motivational problems in most survey interviews, especially in interviews that deal with topics that are embarrassing or otherwise uncomfortable to talk about. The evidence is equally clear that strategies of the sort used in the NCS, including the use of motivational instructions, commitment questions, and contingent feedback, are able to reduce these problems. As a result, most professional academic survey organizations in the US and other Western countries have adopted the use of these strategies. Contingent feedback, in particular, is now a routine part of interviewer training in major academic survey research centres around the world. Yet the WHO CIDI Advisory Committee has not agreed so far to adopt these strategies for use in the CIDI. It is not clear to us why this is the case. Our hope, though, is that the Committee will re-evaluate this decision and include the use of these strategies in future versions of the CIDI.

Facilitating accurate reporting

The main effort in the NCS, to facilitate accurate reporting beyond the use of motivational techniques, focused on lifetime recall of diagnostic stem questions. Our strategy here was to develop the lifetime review section described earlier in this article and to use interviewer administration procedures (reading the questions slowly, instructing the respondents to take their time and think carefully before answering) that were designed to facilitate active memory search. The debriefing of pretest respondents suggests that these procedures did, in fact, lead to more active memory search. This is something that can be especially important among older respondents, as the reconstructive process gets more difficult and the length of the recall period increases. The results of the experiment reported in Table 1 showed clearly that the more active memory work resulted in a significant increase in the proportion of respondents who endorsed lifetime diagnostic stem questions.

Despite this evidence, the WHO CIDI Advisory Committee has not agreed so far to adopt the life review section as part of the official CIDI. The Committee clearly recognizes that the use of a life review section significantly increases the proportion of respondents endorsing stem questions and, in this way, increases lifetime prevalence estimates of disorders. However, their concern, as expressed to us by a number of Committee members, is that this might result in an increase in false positive diagnoses.

This is a legitimate concern in the abstract but it is inconsistent with the facts. The most important fact is that, as reported earlier in this paper, the NCS clinical reappraisal study found no evidence of overestimation of lifetime prevalences in the NCS. It is worth noting in this regard that only a minority of general population respondents who endorse a diagnostic stem question go on to meet full diagnostic criteria for that disorder in the CIDI. Our assumption in developing the life review section was that the large number of symptom questions administered after a positive stem response would be responsible for sorting out true cases from non-cases. The NCS clinical reappraisal study shows this to be the case.

Far from guarding against overdiagnosis, failure to adopt a life review approach to recall has the adverse effect of promoting underdiagnosis of lifetime disorders in cross-sectional general population epidemiologic surveys. This, in turn, creates two important false impressions about the epidemiology of psychiatric disorders: that they are more rare than they really are and that, when they occur, they have a very high probability of becoming chronic or recurrent. In addition, this underdiagnosis in baseline cross-sectional surveys leads to serious difficulties of interpretation in longitudinal follow-up surveys. As we saw above in the discussion of the second wave of the ECA study, underestimation of lifetime prevalences in a baseline survey leads to unrealistically high estimated rates of first onset between the baseline and subsequent waves. These unrealistically high estimates, in turn, make it difficult to study incidence. It is largely due to this difficulty that the

two-wave ECA panel data have never been adequately analysed even though NIMH invested millions of dollars in collecting these data. As noted above, the work of Turner and his colleagues, who carried out a two-wave survey similar to the ECA in Toronto, shows that these unrealistically high incidence estimates are not found when the life review section is used to stimulate lifetime recall in the baseline survey.

Based on these considerations, we believe that the life review section should become a standard part of the CIDI. In addition, we believe that much more should be done to modify the CIDI so as to improve accuracy of recall of answers to other complex question. There are a great many questions in the CIDI that pose difficult memory challenges for respondents, including questions about the age of onset of disorders, the number of lifetime episodes of particular disorders, the length of longest lifetime episodes, and symptom clusters present in particular episodes. The work of cognitive psychologists reviewed earlier in this paper has dealt with questions of these types in other contexts and has shown that it is often possible to improve reporting accuracy by using a variety of memory-enhancing techniques such as decomposition (breaking down complex memory tasks into easier component parts) and anchoring (asking whether particular occurrences happened before or after salient marker events rather than at particular ages). These techniques all take advantage of the knowledge amassed by cognitive scientists over the past decade about the ways in which information is stored in memory and the strategies that are most effective is recovering this information. A serious programme of research in which the WHO CIDI Advisory Committee actively collaborated with cognitive psychologists to study memory processes involved in answering CIDI recall questions would probably result in a number of CIDI modifications that would substantially improve reporting accuracy.

Recognizing the limits of reporting accuracy

The CIDI can be faulted not only because it includes a great many difficult lifetime recall questions without providing memory aids but also because it does not acknowledge the difficulty and, in some cases, the impossibility of providing accurate answers to these questions. As reviewed earlier in the paper, the literature on the limits of memory makes it perfectly clear that failure to acknowledge the limits of memory can lead to reductions in data quality because it encourages guessing, fails to motivate serious memory search within the limits of memory, and provides no means of capturing accurate partial information.

Our main modifications of the CIDI for the NCS concerning the limits of memory focused on questions about age of onset. As described earlier in the paper, we dealt with the memory problem for age of onset by using a three-question series developed by Charles Cannell that began with a meta question ('Can you remember your exact age the very first time . . .?') and then sought to capture partial information ('About how old were you . . .?' and 'What's the earliest age you can clearly remember . . .?') in cases where the respondent did not have a vivid memory of their age of first onset. Our debriefing of pretest respondents convinced us that this sequence is far superior to the standard age of onset question used in the DIS and CIDI, a view confirmed by a subsequent methodological study (Knäuper et al. in press) that documented the disappearance of the apparently biased lumping of age of onset reports five years prior to the time of interview found in the ECA Study (Simon and Von Korff 1992, 1995) in the NCS. Despite this evidence, the WHO CIDI Advisory Committee has not agreed so far to adopt this improved set of age-of-onset questions for reasons that are unclear to us. We believe that these questions should be adopted in future versions of the CIDI and that a series of studies should be undertaken to investigate other ways in which the accuracy of responses to CIDI recall questions can be improved by introducing modifications that recognise the limits of reporting accuracy.

Acknowledgements

The work reported here was carried out in conjunction with the International Consortium in Psychiatric Epidemiology (ICPE). More information about the ICPE can be obtained from http://www.hcp.med.harvard.edu/icpe

Preparation of this chapter was supported, in part, by grants R01 MH41135, R01 MH46376, R01 MH49098, K05 MH00507, and K05 MH01277 (Kendler's RSA) from the National Institute of Mental Health with supplemental support from the National Institute of Drug Abuse (through a supplement to MH46376) and the WT Grant Foundations (Grant 90135190). We are indebted to all the interviewers, field supervisors, and central office staff without whom the NCS could not have been completed. Several of these individuals require special mention. Peggy Price was the field study manager during the pretest phase of the study. Price was responsible for implementing all the three NCS pretests

and worked closely with Kessler to interpret the results and develop CIDI modifications based on the results of the pretests. Barb Hamburg was the field study manager for NCS production interviews. Hamburg worked closely with field staff to maintain high quality control throughout production interviewing. Beth Ellen Pennell was the field director for the NCS. Pennell worked closely with Hamburg and Kessler to design and implement a variety of innovative design and incentive features for interviewers and respondents. Joy Pixley was the field study manager for the national telephone survey experimental evaluation of the CIDI modifications. Pixley single-handedly trained and supervised the interviewers, managed the telephone sample, and built the data file on which analyses of the experimental data were based. Finally, we owe an enormous intellectual debt to Charlie Cannell, who was the methodological consultant for NCS instrument development. Cannell's studies of survey data collection methodology were the foundation on which most of our pretesting procedures and CIDI modifications were based. In addition, Cannell designed the age of onset probe question used in the NCS. A complete list of NCS publications, study documentation, and information on the NCS public use data tape can be obtained from the NCS home page at http://www.hcp.med.harvard.edu.ncs.

References

Belli RF. Color blend retrievals: compromise memories or deliberate compromise responses? Memory and Cognition 1988; 16: 314–26.

Belson WA. The Design and Understanding of Survey Questions. Aldershot, England: Gower, 1981.

Biderman A. Report of a Workshop on Applying Cognitive Psychology to Recall Problems of The National Crime Survey. Washington DC: Bureau of Social Science Research, 1980.

Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis: Theory and Practice. Cambridge MA: MIT Press, 1975.

Blazer DG, Kessler RC, McGonagle KA, Schwarz MS. The prevalences and distribution of major depression in a national community sample: the national comorbidity survey. Am J Psychiatry 1994; 151: 979–86

Bradburn N, Sudman S, Blair E, Locander W, Miles C, Singer E, Stocking C. Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research. San Francisco: Jossey-Bass, 1979.

Bradburn NM, Rips LJ, Shevell SK. Answering autobiographical questions: the impact of memory and inference on surveys. Science 1987; 236: 157–61.

Brewer WF. What is autobiographical memory? In DC Rubin (ed.) Autobiographical Memory. New York: Cambridge University Press, 1986, pp. 25–49.

Bromet EJ, Dunn LO, Connell MM, Dew MA, Schulberg HC. Long-term reliability of diagnosing lifetime major depression in a community sample. Arch Gen Psychiatry 1986; 43: 435–40.

Burton S, Blair E. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. Public Opinion Quarterly 1991; 55: 50–79.

Cannell CF. Experiments in the improvement of response accuracy. In Beed TW, Stimson RJ (eds) Survey Interviewing: Theory and Techniques. Winchester MA: Allen & Unwin, 1985a, pp. 24–62.

Cannell CF. Overview: response bias and interviewer variability in surveys. In Beed TW, Stimson RJ (eds) Survey Interviewing: Theory and Techniques. Winchester MA: Allen & Unwin, 1985b, pp. 1–23.

Cannell CF, Fowler FJ, Marquis KH. The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. Vital Health Stat 1968; 26(2) 1–65.

Cannell CF, Kahn R. Interviewing. In Lindzey G, Aronson E (eds) The Handbook of Social Psychology. Reading MA: Addison-Wesley, 1968, Vol. 2, pp. 526–95.

Cannell CF, Miller PV, Oksenberg L. Research on interviewing techniques. In Leinhardt S (ed.) Sociological Methodology. San Francisco: Jossey-Bass, 1981, pp. 389–437.

Cannell CF, Oksenberg L, Converse JM. Striving for response accuracy: experiments in new interviewing techniques. Journal of Marketing Research 1977; 14(3): 306–15.

Clark HH, Schober MF. Asking questions and influencing answers. In JM Tanur (ed.) Questions About Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation, 1992, pp. 15–48.

Cohen J. A coefficient of agreement for nominal data. Education and Psychological Measurement 1960; 20: 37–46.

DeMaio TJ, Rothgeb JM. Cognitive Interviewing Techniques: In the Lab and in the Field. In Schwarz N, Sudman S (eds) Answering Questions: methodology for determining cognitive and communicative processes in survey research. San Francisco: Jossey-Bass 1996; pp. 177–95.

Ericson KA, Simon HA. Protocol Analysis. Verbal Reports as Data. London, Cambridge MA: MIT Press, 1993.

Fleiss JL. Statistical methods for rates and proportion (2 edn) New York: John Wiley & Sons, 1981.

Fowler FJ Jr, Cannell CF. Using behavioral coding to identify cognitive problems with survey questions. In Schwarz N, Sudman S (eds) Answering Questions: methodology for determining cognitive and communicative processes in survey research. San Francisco: Jossey-Bass, 1996, pp. 177–95.

Genter D, Collins A. Studies of inference from lack of knowledge. Memory and Cognition 1981; 9: 434–43.

Glucksberg S, McCloskey M. Decisions about ignorance: knowing what you don't know. J Exp Psychol Human Learning and Memory 1981; 7: 311–25.

Jabine T, Straf M, Tanur JM, Tourangeau R (eds). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. Washington DC: National Academy Press, 1984.

Jefferson G. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In Roger D, Bull P (eds) Conversation: An Interdisciplinary Perspective. Philadelphia PA: Multilingual Matters, 1989, pp 166–96.

Jobe JB, White AA, Kelley CL, Mingay DL, Sanchez MJ, Loftus EF. Recall strategies and memory for health care visits. The Milbank Quarterly 1990; 68: 171–89.

Kessler RC, Crum RM, Warmer LA, Nelson CB, Schulenberg J, Anthony JC. The lifetime co-occurence of DSM-III-R alcohol abuse and dependence with other psychiatric disorders in the National Comorbidity Survey. Arch Gen Psychiatry 1997; 54: 313–21.

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen HU, Kendler KS. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. Arch Gen Psychiatry 1994; 51: 8–19.

Kessler RC, Mroczek DK, Belli RF. Retrospective adult assessment of childhood psychopathology. In Schaeffer D, Richters J (eds) Assessment in and Adolescent Child Psychopathology. New York: Guilford Press, in press.

Kessler RC, Sonnega A, Bromet E, Hughes M, Nelson CB. Posttraumatic stress disorder in the national comorbidity survey. Arch Gen Psychiatry 1995; 52: 1048–60.

Kessler RC, Wethington E. The reliability of life event reports in a community survey. Psychol Med 1991; 21: 723–38.

Knäuper B, Kessler RC, Cannell CF, Schwarz N, Bruce ML. 'When Was the First Time . . .?' Improving the Reliability and Validity of Age of Onset Reports, in press.

Lessler J, Salter W, Tourangeau R. Questionnaire design in the cognitive research laboratory: results of an experimental prototype. Vital and Health Statistics. Washington DC: US Government Printing Office 1989; 6(1) (DHHS Publication No. PHS 89-1076).

Mannuzza S, Fyer AJ, Martin LY, Gallops MS, Endicott J, Gorman J, Liebowitz MR, Klein DF. Reliability of anxiety assessment. I, Diagnostic agreement. Arch Gen Psychiatry 1989; 46: 1093–101.

Markus H, Zajonc RB. The cognitive perspective in social psychology. In Lindzey G, Aronson E (eds) The Handbook of Social Psychology (3 edn) New York: Random House, 1985; pp 137–230.

Marquis KH, Cannell CF. A Study of Interviewer-Respondent Interaction in the Urban Employment. Survey Research Center. Ann Arbor MI: University of Michigan, 1969.

McLeod JD, Turnbull GE, Kessler RC, Abelson JM. Sources for discrepancy in the comparison of a lay-administered diagnostic instrument with clinical diagnosis. Psychiatric Research 1990; 31: 145–59.

Means B, Loftus EF. When personal history repeats itself: Decomposing memories for recurring events. Applied Cognitive Psychology 1991; 5: 297–318.

Means B, Swan GE, Jobe, JB, Esposito JL. Estimating frequencies for habitual behaviors: reports of cigarette smoking. In Schwartz N, Sudman S (eds) Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag, 1993, pp. 107–20.

Menon A. Judgements of behavioral frequencies: memory search and retrieval strategies. In Schwartz N, Sudman S (eds) Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag, 1994, pp. 161–72.

Miller PV, Cannell CF. Communicating measurement objectives in the survey interview. In Hirsch DM, Miller PV, Kline FG (eds) Strategies for Communication Research. Beverly Hills CA: Sage, 1977, vol. 6.

Moss C, Goldstein H (eds). The Recall Method in Social Surveys. London: NFER Publishing, 1979.

Neter J, Waksberg J. A study of response errors in the expenditures data from household interviews. Journal of the American Statistical Association 1964; 59: 18–55.

Oksenberg L, Cannell CF, Kalton G. New strategies for pretesting survey questions. Journal of Official Statistics 1991; 7: 349–65.

Oksenberg L, Vinokur A, Cannell CF. The effects of instructions, commitment and feedback on reporting in personal interviews. In Cannell CF, Oksenberg L, Converse JM (eds) Experiments in interviewing techniques, Department of Health, Education, and Welfare. Washington, D.C., pp. 133–99 DHEW Publication No. (HRA) 78-3204, 1979a.

Oksenberg L, Vinokur A, Cannell CF. Effects of commitment to being a good respondent on interview performance. In Cannell CF, Oksenberg L, Converse JM (eds) Experiments in Interviewing Techniques. Washington DC: Department of Health, Education, and Welfare, pp. 74–108, DHEW Publication No. (HRA) 78-3204, 1979b.

Pearson RW, Ross M, Dawes RM. Personal recall and the limits of retrospective questions in surveys. In Tanur JM (ed.) Questions about Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation, pp. 65–94, 1992.

Regier DA, Kaelber CT, Rae DS, Farmer ME, Knäuper B, Kessler RC, Norquist GS. Limitations of diagnostic criteria and assessment instruments for mental disorders: Implications for research and policy. Arch Gen Psychiatry, in press.

Robins LN. Epidemiology: Reflections on testing the validity of psychiatric interviews. Arch Gen Psychiatry 1985; 42: 918–24.

Robins LN, Schoenberg SP, Holmes SJ, Ratcliff KS, Benham A, Works J. Early home environment and retrospective

recall: a test for concordance between siblings with and without psychiatric disorders. Am J Orthopsychiatry 1985; 55: 27–41.

Robins LN, Locke BZ, Regier DA. An Overview of Pychiatric Disorders in America. In Robins LN, Regier DA (eds) Psychiatric Disorders in America: The Epidemiologic Catchment Study. New York: Free Press, 1991, pp. 328–66.

Robins LN, Regier DA. Psychiatric Disorders in America: The Epidemiologic Catchment Study. New York: Free Press, 1991.

Schaeffer NC. Hardly ever or constantly? Group comparisons using vague quantifiers. Public Opinion Quarterly 1991; 55(3): 395–423.

Schwarz N. Assessing frequency reports of mundane behaviors: contributions of cognitive psychology to questionnaire construction. In Hendrick C, Clark MS (eds) Research Methods in Personality and Social Psychology. Beverly Hills CA: Sage 1990; 11: 98–119.

Schwarz N, Sudman S. Context Effects in Social and Psychological Research. New York: Springer Verlag, 1992.

Schwarz N, Sudman S. Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer Verlag, 1994.

Schwarz N, Sudman S. Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco CA: Jossey-Bass, 1996.

Shannon B. Have you ever been to Paris? Acta Psychol (Amst) 1979; 43: 313–28.

Simon GE, VonKorff MR. Reevaluation of secular trends in depression rates. Am J Epidemiol 1992; 135: 1411–22.

Simon GE, VonKorff MR. Recall of psychiatric history in cross-sectional surveys: implications for epidemiologic research. Epidemiol Rev 1995; 17: 221–7.

Spitzer RL, Williams JBW, Gibbon M, First MB. The structured clinical interview for DSM-III-R. History, rationale and description. Arch Gen Psychiatry 1992; 49: 624–9.

Sudman S, Bradburn NM. Response Effects in Surveys: A Review and Synthesis. Chicago IL: Aldine, 1974.

Sudman S, Bradburn NM. Asking Questions: A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass, 1982.

Sudman S, Bradburn N, Schwarz N. Thinking about Answers: The Applications of Cognitive Processes to Survey Methodology. San Francisco CA: Jossey-Bass, 1996.

Tanur JM (ed.). Questions About Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation, 1992.

Turner CF, Lessler JT, Grfoerer JC. Survey measurement of drug use. Methodological studies. Washington DC: US Department of Health and Human Services, 1992.

Vinokur A, Oksenberg L, Cannell CF. Effects of feedback and reinforcement on the report of health information. In: Cannell CF, Oksenberg L, Converse JM (eds) Experiments in Interviewing Techniques. University of Michigan, Ann Arbor, MI: Survey Research Center, 1979.

Warner LA, Kessler RC, Hughes M, Anthony JC, Nelson CB. Prevalence and correlates of drug use and dependence in the United States: Results from the National Comorbidity Survey. Arch Gen Psychiatry 1995; 52: 219–29.

Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HG, Rounsaville B, Wittchen HU. The Structured Clinical Interview for DSM-III-R (SCID) II: Multi-site test retest reliability. Arch Gen Psychiatry 1992; 49: 630–6.

Wittchen H-U. Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI). A critical review. J Psychiatr Res 1994; 28: 57–84.

Wittchen H-U, Kessler RC, Zhao S, Abelson JM. Reliability and clinical validity of UM-CIDI DSM-III-R generalized anxiety disorder. J Psychiatr Res 1995; 29: 95–110.

Wittchen H-U, Zhao S, Abelson JM, Abelson JL, Kessler RC. Reliability and Procedural Validity of UM-CIDI DSM-III-R phobic disorders. Psychol Med 1996; 26: 1169–77.

World Health Organization. Composite International Diagnostic Interview (CIDI, Version 1.0). Geneva: World Health Organization, 1990.

Address comments to the first author at the Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02160. 617-432-3587 (voice), 617-432-3588 (fax).