Making a Structured Psychiatric Diagnostic Interview Faithful to the Nomenclature

Lee N. Robins and Linda B. Cottler

From the Department of Psychiatry, Washington University School of Medicine, St. Louis, MO.

Received for publication February 23, 2004; accepted for publication May 19, 2004.

Psychiatric diagnostic interviews to be used in epidemiologic studies by lay interviewers have, since the 1970s, attempted to operationalize existing psychiatric nomenclatures. How to maximize the chances that they do so successfully has not previously been spelled out. In this article, the authors discuss strategies for each of the seven steps involved in writing, updating, or modifying a diagnostic interview and its supporting materials: 1) writing questions that match the nomenclature’s criteria, 2) checking that respondents will be willing and able to answer the questions, 3) choosing a format acceptable to interviewers that maximizes accurate answering and recording of answers, 4) constructing a data entry and cleaning program that highlights errors to be corrected, 5) creating a diagnostic scoring program that matches the nomenclature’s algorithms, 6) developing an interviewer training program that maximizes reliability, and 7) computerizing the interview. For each step, the authors discuss how to identify errors, correct them, and validate the revisions. Although operationalization will never be perfect because of ambiguities in the nomenclature, specifying methods for minimizing divergence from the nomenclature is timely as users modify existing interviews and look forward to updating interviews based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, and the International Classification of Diseases, Eleventh Revision.

cohort studies; data collection; epidemiologic methods; interviews; mental disorders; psychiatry

Abbreviations: CIDI, Composite International Diagnostic Interview; DIS, Diagnostic Interview Schedule.

Beginning in the 1970s, structured psychiatric diagnostic interviews have been used to make specific diagnoses according to a standard nomenclature (1, 2). The Schedule for Affective Disorder and Schizophrenia (SADS) (3), Structured Clinical Interview for DSM-III-R (Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised) (SCID) (4), and Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (5) were designed for administration by a clinician. The Diagnostic Interview Schedule (DIS) (6–9) and the Composite International Diagnostic Interview (CIDI) (10–13) were more fully structured so that lay interviewers could be trained to replace clinician-interviewers; diagnoses were scored by computer, which made the DIS and CIDI appropriate for estimating diagnostic prevalences in large epidemiologic studies.

The validity of the prevalence estimates for mental disorders achieved in these epidemiologic studies is not easy to determine. At best, it cannot be greater than that of the nomenclature the interview serves. Why is it important that diagnoses in epidemiologic studies be faithful to a nomenclature of uncertain validity? The official nomenclatures from 1980 onward have greatly improved communication (14). Epidemiologic diagnostic results, for which interviews faithful to the official nomenclature are used, can be correctly understood by anyone consulting the official diagnostic manual. Otherwise, there is room for endless doubts about whether persons given a positive diagnosis “really” had that disorder. Psychiatry does not yet have convincing ways to recognize “real” disorders; till then, we will have to settle for asking whether the interview successfully identifies the disorders as described in the manual (15).

To prevent a study’s validity from being less than that of the nomenclature’s, its interview must correctly interpret the nomenclature, its questions must be readily understood and acceptable, it must be presented in standard fashion to achieve reliability, and its answers must be recorded correctly. Responses must be scored according to the nomenclature’s diagnostic algorithms.

Reprint requests to Dr. Lee N. Robins, Department of Psychiatry, Washington University School of Medicine, Box 8134, 660 South Euclid Avenue, St. Louis, MO 63110 (e-mail: robinsl@psychiatry.wustl.edu).
Errors can occur at each stage in the construction of an interview. This paper grew out of our long history of writing and revising structured diagnostic interviews (16). We suggest strategies for identifying and correcting errors at each stage and for verifying that the modified versions remain at least as faithful to the nomenclature as the original interview. These strategies should be useful as existing interviews are modified to fit future versions of the nomenclature or as new interviews are constructed. Some, but not all, of these strategies have been used as versions of the DIS and CIDI were tested and modified to match successive editions of the Diagnostic and Statistical Manual of Mental Disorders and International Classification of Diseases and serve cross-cultural studies (10, 17–19).

THE SEVEN STEPS IN INTERVIEW CONSTRUCTION

Writing diagnostic questions

The diagnoses to be made are divided among the authors, who then write questions that follow the manual as faithfully as possible while using language expected to be comprehensible and acceptable to respondents. At least one question is devoted to each diagnostic criterion. These criteria include symptoms, duration of symptoms, age at onset, chronicity, impairment, and overlaps in time between this disorder’s own symptoms and symptoms of possibly preemptive diagnoses.

Symptom questions. Questions must cover symptoms whenever they occurred in the respondent’s lifetime to allow assessing the manual’s criteria for the minimum number of symptoms. They must also ask when symptoms first occurred and last occurred to assess whether criteria for age at onset and duration were met. Questions are also needed for ascertaining in what years symptoms were present, not only for the diagnosis of interest but also for all diagnoses that may preempt it, because the diagnosis will be preempted if its symptoms occurred only when a possibly preemptive disorder was active. Dating of symptoms also allows assessing whether disorders are currently active.

Psychiatric relevance. Many symptoms of psychiatric disorders resemble symptoms of physical diseases, injury, or substance ingestion. For each symptom, the interview must enable a decision as to whether the symptom was plausibly explained by psychiatric disorder. Probe questions are written (and repeated for each symptom) to exclude reported symptoms that either do not qualify as causing impairment or distress or can be fully explained by physical causes (20).

Interpreting nonspecific words. While the diagnostic manuals written in 1980 and thereafter offer vast improvement over earlier versions regarding the specificity with which they describe criteria, they are still not totally explicit. The manual often suggests that there may be relevant symptoms in addition to those it lists. For example, for Specific Phobia, the phobia concerns “a specific object or situation (e.g., flying, heights, animals, receiving an injection, seeing blood)” (21, p. 410); “e.g.”s suggest that other symptoms would also qualify. However, we do not add symptoms when there is an “e.g.” because there would be no official sanction that those chosen are appropriate.

The manuals use terms such as “persistent,” “markedly increased,” “excessive,” “intense,” and “recurrent.” If the interview were to use these terms, subjects would ask the interviewer to be more precise: “Well, what would you call ‘excessive’?” The traditional interviewer’s response, “Whatever it means to you,” is not satisfactory because reliability requires that the word mean the same thing to every respondent. Our solution has been to choose a quantitative equivalent and to use it consistently throughout the interview (22, 23).

Assessing the questions. The first step in assessing the author’s success in writing appropriate questions is to have all other authors review his or her work. These authors consider whether all symptoms have been assessed and whether symptoms are assessed for both lifetime and present occurrence. The authors circulate suggested revisions and then meet to reach consensus on each question.

As they work closely together, the authors may begin to think too much alike and fail to recognize problems with each others’ questions. Once they reach consensus, they should call upon outside experts to review the questions’ appropriateness. Rewriting of questions found to be defective follows this expert review. The revisions are then reviewed by all authors and changes are made until consensus is again reached.

Testing respondents’ reception of the questions

To answer the interview’s questions correctly, respondents must understand them, have the information requested, and be willing to share it with an interviewer.

Respondents’ understanding. The authors’ success in translating criteria into clear, simple language is tested by interviewing small groups of respondents. These persons are chosen to represent a wide spectrum of literacy and social backgrounds.

A question is read to respondents, who are then asked to rephrase the question in their own words and answer it. If the rephrasing means what the authors intended the question to mean, the question’s topic is understandable. To decide whether the answers match the authors’ expectations of what a positive or negative answer should mean, respondents who gave a positive answer are asked to describe their experiences with the symptom—when it occurred, how long it lasted, what it was like. Respondents who gave a negative answer are asked the same questions about any experience they had that was at all similar to the symptom. If the borderline between positive and negative examples does not correspond to the distinction the authors intended, the question needs revision.

Having respondents rephrase questions and describe their symptoms takes considerably more time than would ordinary administration of the final interview. To keep respondents and interviewers fresh and attentive, diagnoses can be divided among several groups of respondents.

Questions to which the respondent knows the answer. The manuals set a minimum frequency and duration for some symptoms, particularly those that often occur transiently in psychiatrically healthy people. Other symptoms count only if they first occur before a specified age. To ask
respondents whether they meet these criteria, it would seem reasonable to ask questions such as, “How often did you … ?” “How long did it last?” “When did you first … ?”. Yet most respondents will not know the answer. They would have to make an estimate of these numbers on the spot. Having to estimate makes responses slow and unreliable and yields a high rate of “don’t know” responses. Frequent “don’t knows” and poor reliability indicate a need for revision (24).

Questions can minimize the precision of recall demanded. For example, “How many panic attacks like that have you had?” can be replaced with “Did you have attacks like that at least four times?” This wording would still make it possible to decide whether the manual’s minimum criterion of four or more attacks had been met. Using quantities specified in the manual reduces the “don’t know” answers and speeds up the interview because respondents often know that the number was far greater than the number meeting the criteria, and they agree rapidly.

Obtaining honest answers. Symptoms of psychiatric disorder that involve sexual behavior, alcohol abuse, and so forth, may embarrass a respondent or be considered too private to discuss with a stranger. Questions not acceptable to respondents lead to denial of their symptoms or refusal to answer.

Questions likely to lead to dishonesty can be identified by signs of discomfort in respondents answering them and by asking respondents which questions, if any, made them uncomfortable. Such questions can be rephrased to make them less objectionable, or can be preceded by reassurance about confidentiality, or can be put in an audiotape or a questionnaire so that the respondent need not answer the interviewer face-to-face (18).

Testing revisions. After revisions have been made to questions that were misunderstood, that asked for information not readily available to the respondent, or that made the respondent uncomfortable, the revised questions must pass two tests: 1) a similar, new group of respondents must demonstrate that they can answer them easily and correctly; and 2) a comparison with the manual’s text must show that they still correspond closely to the manual’s criteria. Questions that fail either test must be rewritten and retested until success is achieved.

Selecting the format

In this section, we discuss formats for a paper-and-pencil version of the interview, with questions to be read as written and acceptable answers assigned either a code to be circled or a number to be inserted in a blank. As noted above, a questionnaire format may be used for brief sections that the respondents find embarrassing, but questionnaires cannot serve as the principal format because they put too great a burden on the respondent. A computerized version is feasible, but, as we will see later, it should be based on a well-tested paper-and-pencil version of the interview.

Labeled questions. A label for each question in the left margin is a format developed for the DIS and CIDI that has proven very useful. The label shows which nomenclature, which diagnosis, and which criterion of that diagnosis the question serves. Identifying these three levels facilitates reviewing the question’s appropriateness and greatly helps the programmer when constructing the scoring program.

Labels can be compact. As an example, in the CIDI, we gave the label PAN10A to question D56: “Have you more than once had an attack like that that was totally unexpected?” (8). “PAN” meant that the question applied to Panic Disorder; “10” meant that it served the International Classification of Diseases, Tenth Revision; and “A” meant that it served Panic Disorder’s Criterion A.

Labels allow testing as to whether there are missing or unnecessary questions. A criterion in the manual for which there is no matching label shows that a needed question is missing or mislabeled. Unnecessary questions are discovered when they cannot be labeled with a specific criterion. Redundancy may be suspected when two or more questions have the same label, although some criteria do indeed require multiple questions.

To verify that all labels needed are present and correct, the label-question pairs are sorted alphabetically by the label field. An author looking at a criterion in the manual says aloud what the label of that criterion should be but does not read the criterion aloud. An assistant searches for that label on the alphabetic list. If it is found, he or she reads its associated question(s) aloud. If the author looking at the diagnostic manual judges that a positive answer to the question(s) would satisfy the criterion, the assistant checks off the label. If the label is not found, the criterion is marked, showing either that there is no question to cover it or that the appropriate question was mislabeled. This exercise is repeated until all criteria for each diagnosis are considered. At the end, the question associated with each unchecked label is reviewed to see whether it should be relabeled to correspond to a marked criterion or whether it is unnecessary and could be deleted. Questions are added to cover marked criteria for which no mislabeled question yet exists.

Disputed formatting issues. Uncertainty exists about which other formats cope best with the complexities intrinsic to diagnostic interviews because there have been few studies of the consequences of adopting one format versus another. An exception is work on revising the CIDI (11). Yet it remains difficult to defend any particular choices. We describe here some of the decisions that must be made and studies that could guide the authors’ decisions.

Screener versus simple modular structure. The older interviews placed each diagnosis in a separate module. Modular construction allows the researcher to easily shorten the interview by dropping the modules for diagnoses in which he or she is not interested. Another option is to begin with a screener, that is, a series of one or two critical symptom questions for each diagnosis (13). Negative screener answers indicate that that diagnosis’s module should be skipped when it appears later.

The effect of using a screener is not obvious. It certainly saves time because it allows the interviewer to skip questions in the modules for which the screener was negative. However, it produces false negatives for any respondents who screen negative but would have reported enough symptoms in the strictly modular version to meet the criteria. It produces false positives for any respondents positive for the
screener who feel obliged to justify their positive answers to the screener by exaggerating symptoms asked about later.

Checklists versus review of previous responses. The DIS and CIDI both require the interviewer to refresh the respondent’s memory about his or her positive answers to a syndrome’s symptom questions when asking for age at first and last symptom, clustering of symptoms, and comorbidity with other disorders. As an alternative to the interviewer’s riffling through previous answers to recapitulate the positives, he or she may be given a checklist on which to check off each positive symptom after coding it on the interview form. The interviewer then refers only to the checklist when recapitulating. It is not known whether checklists reduce or increase interviewer error. Is recapitulation more complete because the interviewer would have missed some positive symptoms when thumbing through completed pages, or are positive symptoms often missed because the interviewer failed to check them off?

A probe flow chart versus imbedded probes. Probe questions are used to evaluate a symptom’s clinical significance and probable psychiatric relevance. These questions are repeated for almost every symptom. They can be imbedded into the printed interview after each symptom, or they can be listed in generic form in a probe flow chart that instructs interviewers to insert the particular symptom being discussed and to continue along the path specified by the coding options shown on the interview form. The probe flow chart format greatly reduces the interview form’s bulk and has been shown to work quite well (20). However, it is not known how often interviewers omit probes or ask them incorrectly because they fail to consult the chart.

Questionnaires and audiotapes. Questions thought to be embarrassing can be put on audiotapes or into questionnaires to give respondents privacy in responding to them. This strategy has been found to produce more positive answers. Is that because greater privacy leads to greater honesty, or is the higher rate of positive answers explained by random errors caused by the respondent’s mishearing the tape, misreading the questionnaire, or accidentally circling the wrong answer on the questionnaire? Random error inevitably increases the apparent prevalence if the symptom is actually rare (25).

Coding missing data. For each question, several codes are available to explain why a question was not answered: the respondent replied “I don’t know,” the respondent refused to answer the question, or the interviewer accidentally failed to ask it. Interviewers are told what these codes are, but often the codes are not printed on the interview form. The rationale for their omission is that their presence would tempt interviewers to make less effort to get substantive answers. Does omitting them have this effect, or does their absence lead the interviewer to circle a printed code even when the correctness of that answer was by no means clear?

Studies to resolve these choices. Studies could be undertaken to decide which of these formatting alternatives produces the more complete and accurate information. Two interviewers, each using one of the two alternative formats, would both interview a group of respondents. The respondent would then be asked to explain any discrepancies between his or her answers and to say which was correct. The format producing more accurate answers for the majority of respondents would be selected. If the assets and disadvantages of the alternative formats allow no clear choice between them, interviewers would be asked which format they prefer, and that format would be adopted.

Constructing a program to enter responses into the computer

Once the format has been decided, a computer program for data entry and cleaning is constructed to enter interview responses into a computerized data set, ready for analysis.

Responses are entered in question order, but the program stops for “cleaning” when an entry is not logically consistent with a previous entry (e.g., the age at remission is lower than the age at onset) or when an answer is expected for which nothing has been coded on the interview. Once the error has been corrected, data entry continues.

Four explanations are possible if the data entry program stops for cleaning: the data entry program is in error, the interview form has incorrect skip instructions, the interviewer failed to ask a required question or coded its answer incorrectly, or the data entry clerk made a keying-in error. Another indication of error is if the data entry program does not stop to ask for entry of a code circled in the interview. There are three possible explanations: a missing skip instruction in the interview, an unnecessary skip instruction in the data entry program, or a failure by the interviewer to follow the interview’s skip instructions.

Thus, as the data entry program is used, errors are discovered simultaneously in that program and in the interview format. The editor reviews both the interview and the data entry program to decide whether either is the source of the problem. If so, the data entry program or the interview form must be corrected.

Devising a scoring program to make diagnoses

The scoring program evaluates each diagnostic criterion and then combines all of them to make diagnoses according to the manual’s algorithms. For each respondent, each diagnosis is scored as present, positive criteria met but possible preemption, negative, or insufficient information to be sure it is negative (26). The score is added to the data set, and a report is prepared of the respondent’s results.

Errors in the program will generally come to light as the program is used, but it would be valuable to be able to correct them prior to use. Finding no advice in the literature on how to conduct a formal test of scoring programs, we devised a method to do so (27): Two programmers independently constructed a scoring program. Then, the computer created a large pseudo-data set that obeyed all of the interview’s rules for answering or skipping questions by randomly assigning one of the logically possible codes to each question to be answered. Every pseudo-case was scored with both programs. Each disagreement between their results meant an error in one or the other program, and the program with the error was corrected. The process was then repeated until the two programs agreed on the presence or absence of each criterion and each diagnosis for all of the computer-
A computerized interview can provide a diagnostic report immediately after the interview is complete. It can also be designed to offer researchers a variety of options: to omit some disorders, to report on only those disorders currently or recently active, or to use an abbreviated version for some or all disorders. Each of these options has been offered by one or more of the computerized interviews constructed more recently (9, 28–30).

Errors in the computerized interview can be located by entering it a set of completed and cleaned interviews obtained by using the final paper-and-pencil version. If the computer accepts each of the coded answers from the paper-and-pencil interviews, does not ask for answers where none appear in the paper-and-pencil version, and produces the identical diagnoses, the computerized version is validated. Otherwise, the source of the error must be located and corrected.

**DISCUSSION**

We have reviewed each step needed to create or modify a fully standardized diagnostic interview that produces diagnoses faithful to the official nomenclature. At each stage, errors may occur. We have suggested methods for discovering them, correcting them, and validating the corrections. How the interview interprets the nomenclature will still be somewhat uncertain because the nomenclature is not always fully explicit, and errors may slip through despite the cautions and testing suggested here. Still, the resulting interview should come close to making diagnoses according to the nomenclature’s specifications.

This article has not recommended the traditional test for validity—having a study’s respondents reinterviewed by a clinician. There are two problems with that test. First, it provides only an up-or-down vote. It does not show the authors where problems lie or how to correct them. Second, even if the interview’s diagnoses agree with the clinician’s, we cannot know whether the clinician’s diagnoses were faithful to the manual (15, 25). If they were not, the interview’s diagnostic results will not be understood by interested persons who were not party to how the interview was constructed.

The thorough evaluation this article recommends may seem daunting. However, carrying out any portion of these evaluations and making revisions accordingly should improve the correspondence between a new or revised interview and the nomenclature it attempts to implement.

**ACKNOWLEDGMENTS**

The authors gratefully acknowledge the careful reading and helpful suggestions made by Dr. Arbi Ben Abdallah and support from the National Institute of Mental Health (MH17104) and National Institute on Drug Abuse (DA007313).

**REFERENCES**

1. Robins LN. The development and characteristics of the NIMH...


