A critical component of establishing the overall relationship between environmental stressors and individual health and well-being is the utilization of a comprehensive, reliable, and valid measure of stressful life events and difficulties. Two methods of life events measurement predominate: checklist measures (see Chapter 2, this volume) and intensive personal interview measures, which use qualitative probes in order to specify more precisely the characteristics of life events believed to produce stress. These two sorts of measures have evolved from different, although not mutually exclusive, theories of what constitutes stress. As a concomitant to their origins in different theories of stress, these measures have also been shaped in response to different research questions and paradigms.

Checklist methods were informed by a theoretical perspective on stress which asserts that the magnitude of social and environmental change bringing about a need for readjustment is the basis of experienced stress (see Chapter 2, this volume, for a review). The early development of personal interview methods that use qualitative probes, however, was driven by a perspective which assumed that social and environmental changes (and anticipations of those changes) which threaten the most strongly held emotional commitments are the basis for experienced severe stress, and that it is severe stress which threatens health (Gorman & Brown, 1992). Other interview measures have since been developed that are based on the change-readjustment paradigm (e.g., Dohrenwend, Raphael, Schwartz, Stueve, & Skodol, 1993); these are all distinguished from the checklists, however, by their use of qualitative probes.

There are other critical differences between checklist and personal interview measures of life events. The life change-readjustment paradigm assumes that positive as well as negative changes are related to the experience of emotional stress; interview methods, however, consistent with their focus on the most emotionally arousing stressors, are designed to elicit reports of negative events. Checklist methods yield a summary score of the stressfulness of changes experienced over a period of time (usually a year), whereas interview methods are designed to elicit reports of specific events that may have triggered an onset of a severe physical or mental illness. In regard to this latter point, checklist measures tend to (but not solely) use...
of multiple and effective illness-for examining illness, the relationship between experiencing stress using continuous physical illness—such as depression or gastrointestinal attacks.

The aim of this chapter is to acquaint readers with the circumstances in which the use of personal interview methods is appropriate; the relative strengths and weaknesses of available personal interview methods; and the logistics and costs of fielding an interview study. It is not the intent of this chapter to argue that personal interview methods are necessary in all cases for measuring exposure to stressors. Personal interview methods are not efficient for correlational or prospective studies, where the purpose is to establish that there is a relationship between environmental stress and a general measure of symptoms. In such instances, the greater expense of using interview measures is not justified. However, there are instances in which the use of personal interview methods is more suited to a particular research question.

### Rationale for the Personal Interview Versus Checklist Approach

The use of the intensive personal interview technique is particularly critical when the relative timing of exposure to a stressor and the onset of an illness is necessary to address a research question. For example, if an investigator is interested in the relationship between experiencing a stressor and the subsequent onset of panic disorder, then both the timing of the events and the onset of panic must be carefully (and independently) dated. Precise dating is necessary not only to establish the relationship of stress exposure to onset, but also to identify the aspects of the stressful experience that affect onset. In general, self-administered checklists cannot equal the precision of an interviewer trained to elicit aspects of events critical to examine these research questions (Cannell, Miller, & Oksenberg, 1981). Studies of illness onset, course, and recovery often require such precise dating of event and illness onset/offset.

Checklist methods also seem to be more prone to “telescoping,” or the misdating of distant events into a more recent time period (McQuaid et al., 1992; Raphael, Cloitre, & Dohrenwend, 1991). Checklist methods, moreover, are probably not as effective as more intensive personal interviewing in communicating the importance of accurate answers to the respondents (Cannell et al., 1981). In addition, the face-to-face interview with qualitative probes facilitates the use of memory aids to improve event recall and accurate dating (such as calendars, visual representations of important events of the preceding year, and well-timed reminders of personally salient dates). Such aids have been shown to provide substantially better information than can be collected through relatively unassisted self-report (Sobell, Tonetta, Sobell, Schuller, & Maxwell, 1990). This more accurate timing of event onset and offset, moreover, makes it possible to distinguish and date a series of related events and difficulties, which may be required in a study addressing whether exposure to multiple events is more related to onset than exposure to a single event (Brown & Harris, 1978; McQuaid et al., 1992; Raphael et al., 1991).

Interview measures of life events, however, are most widely promoted for their

Studies of responses to checklist measures have found that many respondents report relatively minor events in response to questions that are designed to elicit only severely stressful events (Dohrenwend, Link, Kern, Shroul, & Markowitz, 1990), such as reporting the death of someone relatively distant as the death of someone "close"; or reporting a broken toe as a serious injury, even when it had no impact on one's life besides short-term pain and inconvenience. In experiments with interview measures of life events, Kessler and Wethington (1991) have found that many people report objectively trivial events in response to questions that were thought to emphasize more than adequately the serious nature of the stressful events that were sought. Reporting minor events as severe, moreover, may be related to the respondent's health status at the time of interview (Bebbington, 1986): Distressed people are more likely to remember negative events than happy people (Fiedler & Stroehm, 1986).

Interview measures of life events, moreover, use a different system for rating the severity of life events. Checklist measures have generally (although not universally) applied "normative life change" or "readjustment" weights—a numerical value reassigned to every event reported or the sum of the number of events experienced over a given period (see Chapter 2). If the event description is relatively more inclusive (e.g., "serious illness"), the life change unit weight may be a poor fit to the severity of the actual event prompting the report. Specifically, a recurrence of cancer arguably has a different impact than a bout of influenza, both of which could prompt a yes answer to "serious illness." Studies comparing checklist to interview methods have confirmed that variability of this sort occurs when subjects respond to checklists. Raphael et al. (1991) and McQuaid et al. (1992) demonstrated that respondents to checklists report experiences to life event questions that do not match the researcher's intent. Furthermore, respondents are known to report minor or even positive events in response to questions that were designed to elicit only highly negative and undesirable events (McQuaid et al., 1992).

Event list comprehensiveness has also been a critical issue in the debate between the use of checklist and personal interview methods. Self-report event lists have been criticized for lacking coverage of singular and rare events that may be among the most distressing (Brown, 1981), although more comprehensive checklist methods have been developed in response to this criticism (Zimmerman, 1983). Personal interview methods allow for more flexibility in event elicitation and recording, as well as self-nomination of stressors not evoked by direct questions about domain-specific events and transitions (Brown & Harris, 1978; Dohrenwend et al., 1990).

Interview measures of life events also allow for more effective probing of event reports, an important means of reducing measurement error. An interviewer trained in study procedures can probe to determine if the event reported indeed "matches" the intent of the question and should be counted as a stressor of that type (McQuaid et al., 1992; Raphael et al. 1991). In addition, probing of event responses also allows investigators to determine whether events are connected to one another or whether the same event is mentioned more than once in response to separate questions (Kessler & Wethington, 1991; McQuaid et al., 1992). The latter is useful in order to counter the tendency of some respondents to "overreport" their exposure by mentioning the same event more than once.
critical issues in stress research, primarily because trained interviewers can elicit more detailed descriptive information about variations in event severity over time, and about linkages among related events and difficulties. Interview measures have been used to explore (1) the relationship among alleviation of environmental stressors, illness course, duration, and recovery (e.g., Brown, Adler, & Bisfulco, 1988); (2) the sorts of stressors that are associated with different types of disease onset (e.g., Bebbington et al., 1988; Harris, 1991); (3) and the interaction of individual vulnerability and life history with recent stressors to produce onset of illness (Brown, Bisfulco, & Harris, 1987).

Applications of Interview Measures of Life Events

Interview measures of life events are extremely expensive in comparison to checklist methods. Given their expense, the use of interview methods for measuring life events is most appropriate for field studies or retrospective or longitudinal case-control studies where the relationship of stress exposure to the onset of very serious illness is being investigated (e.g., Katschnig, 1986), and where exposure to life events is of primary interest to the researcher (Cooke, 1985). In such studies, the major research questions entail establishing the relationship of environmental stressors to mental or physical illness in the community. In these studies as well, a maximally comprehensive measure of stress exposure is necessary. The reason is that the focus in such studies is not on one particular event or transition that many people experience, and explaining variations in people’s responsiveness to it, but on establishing the etiology of particular disorders or illnesses, particularly the role of stressful events.

Although personal interviewing methods have also been used extensively in studies of adjustment to one stressor, such as divorce, unexpected unemployment, or widowhood, this chapter will not review such studies, because our emphasis here is on comprehensive measures of stressor exposure. The intention is not to urge, even inadvertently, that researchers abandon intensive studies of one crisis in favor of a comprehensive stressor measurement approach. In a specific crisis study, the comprehensive stressor measures described here would be poor substitutes for the sort of fine-grained examinations of individual variation and adaptation that are typical in such studies.

Choosing an Appropriate Measure

The following section critically evaluates a number of interview measures of life events. The review is confined to those measures that have been used more than once, or that are readily available for replication by researchers other than their developers. The criteria for evaluation include (1) reliability (if reported); (2) comprehensiveness of event coverage; (3) falloff in reports of events over a 1-year event recall period; (4) the relationship of stressor exposure, as estimated by this interview method, to onset of illness; (5) the dimensions by which stressors are rated in order to evaluate their “stressfulness”; (6) illnesses and disorders to which the measure has been applied; (7) populations in which the measure has been used, including cultural groups; (8) strengths of the method and its potential for future development; and (9) major criticisms that have emerged. For several of the reviewed measures, not all of this information is available.

Interview methods are presented as follows: Because of its earlier development, the Brown and Harris Life Events and Difficulties Schedule (LEDS) is described first. It is followed by reviews of a number of interview life-event rating measures that do not reproduce the event rating system used in the LEDS (e.g., the SEPRATE: see below and Dohrenwend et al., 1993). Finally, interview measures that are shortened or more structured versions of the LEDS are reviewed, including a structured form of the LEDS under development (the SLI: see below and Wethington, Kessler, & Brown, 1993) and interviews that use checklists to reduce interviewing time.

The Life Events and Difficulties Schedule

The most widely used personal interview method is the Life Events and Difficulties Schedule (LEDS: Brown & Harris, 1978). The LEDS is a semi-structured survey instrument, appropriate for use in a community sample as well as with patients, assessing a wide variety of stressors. The interview consists of a series of questions asking whether certain types of events had occurred over the past 12 months (or longer) and a set of guidelines for probing positive responses. Interviewers are given a great deal of latitude in deciding whether more intense probing is required for ascertaining whether an event is likely to have been severe in emotional impact, or not. To that end, only general guidelines (brief reminders about important objective dimensions) are given for probing, since the interview is designed to resemble conversation rather than a standard survey.

The aim of the interviewing method is to construct a narrative or “story” of each event (Gorman & Brown, 1992). Once an event is mentioned, the interviewer asks questions about objective circumstances surrounding the event, such as: What led up to the event? What followed it? Interviewers also follow up on spontaneous comments made by the respondents, including mentions of other events occurring at the same time, or what may be a technically “wrong” answer—for example, a parent’s sudden death reported in response to a question about serious illness of a family member.

The purpose of the probes is to gather enough information to rate the long-term contextual threat of discrete events (a death) or severity of chronic difficulties (several months of unemployment). The rating of long-term contextual threat is the key component of the LEDS method, as the experience of a “severely” threatening event or presence of a severe chronic difficulty (“marked” in LEDS terminology) is hypothesized to pose a risk for the onset of physical or mental illness. Rating the degree of threat for specific sorts of events and difficulties has been developed and documented over the several decades in which the LEDS has been used. The documentation for the ratings is contained in dictionaries produced by Brown.
Harris, and their colleagues. These dictionaries, which are made available to those who are formally trained in the use of the method, consist of several thousand examples of rated life events and difficulties, grouped by life domain (marital, work-related, financial), type of event or difficulty (argument, separation, ongoing marital problems), and level of rated severity (4 categories for events, and 7 for difficulties). In addition to the examples, the dictionaries also document how events are classified by type and the sorts of information the interviewer needs to establish in order to assign contextual threat ratings to each type of event.

The critical element of rating long-term contextual threat is collecting objective information in order to estimate the long-term implications of an event or difficulty for important plans, concerns, and purposes of the respondent, taking into account other current circumstances and relevant life history information. Aspects of life history and context that are relevant are documented in the life event and difficulty dictionaries. The general rule guiding event severity ratings is that the interviewer must determine whether the event affected the life plans or commitments of the respondent, and use the dictionary guidelines to assess those aspects of context that must be established in order to estimate the level of emotional arousal likely to have resulted. In many instances, however, the interviewer can rely on the event or difficulty having a specific rating in the LEDS dictionaries.

The event list in the LEDS consists not only of questions about probable negative events, but also of role transitions (births, promotions, graduation from school) and situations that may have involved a severe emotional reaction, but are not typically defined as events in checklists. An unusual feature of the LEDS are questions such as dealing with important revelations about other people’s character, breaking or getting bad news, making difficult decisions, disappointments, and finding presumptive negative events working out much better than expected.

There are three main stages of life-event rating: (1) assessment of the likely emotional arousal (positive or negative) associated with the reported event; (2) a rating of general contextual threat, which has a very high threshold for the highest rating of severity; and (3) ratings of specific aspects of threat, such as its degree of loss (of a person, object, or idea), danger (the future probability of loss), frustration of goals, challenge, intrusiveness into daily routine, and positiveness (e.g., relief, or getting a fresh start on life). Events are rated for both their positive and negative aspects. Difficulties are rated for their intrusiveness into daily routine and their threat to the quality of other life roles, such as work.

The interviews average about one hour to complete in a general population sample, with a range from 30 minutes to well over 2 hours. After the interviewer completes the interview, he or she then constructs written narrative descriptions of each event or difficulty. Interviewer are trained to edit out the respondent’s subjective reactions and emotional responses to the event, so that they cannot confound the event contextual rating. The event descriptions are then presented to a panel of raters, blind to the illness status of the respondent, who discuss the appropriate rating to assign the event or difficulty. Although in the earlier years of LEDS development, most events and difficulties were brought to these panels for discussion and rating, the availability of the detailed event and difficulty dictionaries ratings has routinized a significant proportion of the panel rating.

In addition to long-term contextual threat, a number of other dimensions are coded and rated for each event. These dimensions include type of event (e.g., illness, change in interaction); the person whom the event primarily affected or “focus” (including a “joint” category for events that happened to another but that are likely to have had an emotional impact on the respondent); the amount of threat reported by the respondent; and “independence,” which is the probability that the event or difficulty could have been brought about by erratic behavior caused by the respondent’s illness. The LEDS dictionaries also document the rules for determining these ratings.

The LEDS appears to be very reliable. One-year falloff in event reporting for severe events is about 1 percent per month (Brown & Harris, 1982), although the falloff for noneverse and minor events is notable. Because the occurrence of a severe event (or presence of an ongoing severe difficulty) is found to be related to illness onset in their community studies (e.g., the estimated odds ratio ranges from 7 to 13 over a number of studies of depression; see Paykel, 1978; Surtees & Duffy, 1989), these investigators have argued that their method is more reliable than checklist measurement methods (e.g., Funch & Marshall, 1984).

The LEDS is an evolving instrument, with numerous studies now in progress to expand its reach to populations and illness onsets other than those that were the focus of its early development (schizophrenic patients and British working-class women). The LEDS has subsequently been used with adolescents (Monck & Dobbs, 1985) and the American elderly, with a new LEDS dictionary being developed for the latter (see note 1). A number of LEDS studies have been conducted in racial minority groups in England and Africa (for a review of the cross-cultural studies, see Brown & Harris, 1989b). Neither the LEDS nor LEDS-like methods, however, have been used with children. In addition to studies of schizophrenia and onsets of depression, the LEDS has been used in studies of anxiety, appendectomy, abdominal pain, menstrual disorders, multiple sclerosis, myocardial infarction, and dysphoria (for a review, see Brown & Harris, 1989a).

The LEDS has experienced criticism for its rating and interview methods. The most frequent and persistent criticism is that too much “context” is included in the ratings of contextual threat. Specifically, there is a long-standing controversy over whether its ratings of the magnitude of contextual threat muddle the distinction between the magnitude of event severity and the individual’s “social” vulnerability to a stressor (Dohrenwend et al., 1993; Tennant, Bebbington & Hurry, 1981), thereby inflating estimates of event impact. An example makes this criticism clear. A death event can vary a great deal in severity in the LEDS rating scheme, from “below threshold” for classification as an event to the highest category of severity. The severity rating depends, in part, on the reliance the person placed in the deceased as a confidant, and the number of alternative confidants available. Severity ratings for deaths also depend on other contextual factors that make a death more or less emotionally meaningful, such as closeness of kinship, quality of the predeath relationship, and frequency of contact the years before death (Brown, 1989). However, the inclusion of a social network vulnerability factor in the rating does confound concepts that some researchers may wish to keep separate (Tennant & Bebbington, 1978). The majority of event and difficulty ratings are less influenced by factors related to social vulnerability than is the death example; for instance, ratings
of health events are based on objective contextual factors such as the life threat or disability posed by the illness, and its impact on work and other life routines (Brown, 1989). Similarly, ratings of work losses are based not on social vulnerability but rather on objective factors such as the unemployment rate, supply of jobs comparable to the one lost, and the proportion of family income lost.

Brown and his associates have continued to elaborate their measures of social vulnerability risk factors over time, and to a degree have met some of the criticism regarding the level of confounding between event severity and social vulnerability (Bebbington, 1986). This has been done through the development of separate interview measures of important social vulnerability factors such as commitment to role domains where events may arise, self-esteem, quality of social support, and conflict among social roles (Self-Evaluation and Social Support Schedule: Brown & Harris, 1989). Nor, it should be noted, are all social vulnerabilities included in event ratings. Some aspects of social vulnerability, such as separation from a parent in childhood, are not included in event and difficulty ratings, but are measured and analyzed as additional risk factors for the onset of illness.

The LEDS has not come into general use, and has only recently been used in the United States. A number of European and American researchers have developed alternative instruments; some of them are structured-question elaborations of the life events checklist with contextual probes (e.g., Kessler & Wethington, 1991; Paykel, 1983; Wittchen, Essau, Hecht, Teder, & Pfister, 1989), whereas others are elaborations of the magnitude of "life change" rating systems (Dohrenwend et al., 1993). Other researchers have fielded shortened or modified versions of the LEDS (e.g., Costello & Devins, 1988; Miller et al., 1989; Wethington, Kessler, & Brown, 1993).

There are two explanations for why researchers have sought alternatives to the LEDS. The first is that many researchers are deterred by the cost and complexity of using the semi-structured LEDS interview. Important questions remain about the logistics of combining LEDS methods with the standardized methods of community epidemiological research, including the required 8 days of interviewer training in London or Pittsburgh (see note 1), the complicated rating scheme, and periodic retraining necessary in order to maintain consistent quality of interviews. Respondent burden is also an issue. Interviews are much lengthier than those required for checklist measures of life events (Katschnig, 1986). They also require a level of self-disclosure some respondents and interviewers find uncomfortable. The expense of the panel rating method, moreover, has been viewed as a significant deterrent.

The second reason is that some researchers prefer to make a stricter distinction between magnitude of change brought about by the event and vulnerability, an aim more compatible with the change-readjustment paradigm underlying checklist research methods (see Chapter 2). Recently, moreover, a few investigators have reported that the contextual rating system used in the LEDS may be no more predictive of illness onset that a carefully applied life-change unit rating scheme applied to events reported in a face-to-face interview (e.g., Faravelli & Ambonetti, 1983; Katschnig, 1980, 1986). In addition, within-respondent comparisons of events reported to life-change measures and to a LEDS interview suggest that the two methods yield different event reports (Katschnig, 1980; McQuaid et al., 1992), both of which could be related to onset of illness (Katschnig, 1980). At the time of this review, these questions remain to be more thoroughly investigated.

Alternative Interview and Rating Systems for Life Events

The Standardized Event Rating System

The Standardized Event Rating System (SEPRATE: Dohrenwend et al., 1993) is an alternative life-event interview and rating system, using a magnitude of "life change" rating system.² The SEPRATE, derived from the PERI life-events checklist (Dohrenwend, Krasnoff, Askenasy, & Dohrenwend, 1978) consists of a series of yes/no questions regarding 84 types of events or difficulties that may have occurred and been severely stressful. Each "yes" answer is probed with a number of structured questions, designed to fit a variety of circumstances. Although probes are provided, interviewers are not confined to the exact wordings as written. The purpose of the probes is twofold: to provide a narrative description of each event or difficulty, and to produce a standardized assessment of separate aspects of situations thought to produce experienced stress. These assessments include magnitude of change brought about by the event, desirability of the event (from the average, or "normative" perspective), disruptiveness to daily life routines, threat to life, independence from the respondent's own (possibly) disordered behavior, and "fatefulness" (occurring independently of the respondent's behavior and beyond his or her control to avert) (Dohrenwend et al., 1990).

After the interview is completed, an editor composes a narrative description of each event or difficulty reported. These descriptions differ from those done using the LEDS method, in that they are stripped of any material that might be used to infer "social vulnerability" to especially severe effects from the given stressor. This is done in order to prevent this sort of vulnerability from being inadvertently included in the rating of the event (Dohrenwend et al., 1990). Each event is then rated independently by two raters on the dimensions listed previously (Shrout et al., 1989). It is important to note that not all event context is stripped from the narrative. Indeed, many of the contextual factors taken account of in the LEDS are reproduced by the SEPRATE, such as the life-threatening or disabling aspects of an illness, the length of an unemployment episode, and the objective likelihood of being able to replace a lost job with one paying as well (Dohrenwend, 1993, personal communication).

Dohrenwend and associates have conducted a reliability study comparing the SEPRATE to their previously developed PERI normative life-event rating checklist (Shrout et al., 1989). Falloff for severe events is relatively modest (Raphael et al., 1991). The relative risk of onset of depression associated with "fateful" life events occurring in the previous 12 months is 1.50, and 3.38 for events determined to be both "fateful and highly disruptive" to life routines. The odds ratio associated with fateful and disruptive events is approximately double that estimated by use of the PERI checklist method in the same study (Shrout et al., 1989).

Both the rate of falloff and the case-control odds ratio estimated for onset appear
somewhat unfavorable compared to the LEDS. However, it should be noted that the estimated odds-ratio provided by Dohrenwend and associates may have been driven down by including events and onsets reported to have occurred nearly 3 years before the interview, a longer than usual recall period. Indeed, a small case-control study (less than 25 subjects) fielded by Mazure and Bruce (M. Bruce, 1993, personal communication) indicates that the estimate of relative risk using the SEPRATE is comparable to the LEDS over a recall period more typical of studies of illness onset. It is also the case that rating dimensions are not as fine-tuned in the SEPRATE (see below), perhaps leading to underrating of the potentially severe events. There is also the possibility that the SEPRATE is measuring a set of stressful experiences different from that which the LEDS typically evokes (see Katschnig, 1980). Theoretical refinement of event rating, combined with a shorter recall period, has produced odd-ratio estimates comparable to the LEDS (Dohrenwend, 1994, personal communication).

The SEPRATE is somewhat less comprehensive than the LEDS. The list of events and difficulties is less extensive in it as well as in other interview methods, reviewed below, that have been developed more directly from the LEDS. In particular, questions regarding important revelations about other people’s character, bad news, decisions, and disappointments—an innovative feature in the LEDS—are not included. Questions such as these yield a number of severe events (Brown & Harris, 1978).

Like the LEDS, the SEPRATE technique has been successfully adapted for use in a different culture (Dohrenwend, 1993, personal communication). It has been used in Israel, where a different list of events and difficulties was utilized, although the rating dimensions remained the same. (These different events and difficulties reflected Israeli experience with war and terrorist activities.) The SEPRATE has been used primarily with adult community samples and with patients.

It is unclear whether the use of SEPRATE interview may result in cost savings compared to the LEDS; however, not all information is available for comparison. In one study using the SEPRATE, interview length ranged from 60 to 90 minutes (Lennon, Dohrenwend, Zautra, & Marbach, 1990). Time estimates for narrative production, coding, and rating, however, are not available on a per interview basis. Training in the method is extensive, and (like the LEDS) effective use would require that investigators consult with the developers of the SEPRATE (Dohrenwend, 1993, personal communication).

Other Interview Measures

Other qualitative probe elaborations of event checklists have also been fielded by a number of investigators. These systems include the Detroit Couples Study Life Events Method (Kessler & Wethington, 1991); the Munich Events List (Wittchen et al., 1989); the Paykel Life Events Interview (Cooke, 1985; Paykel, 1983); and the Henderson, Byrne, and Duncan-Jones (1981) List of Recent Experiences. There are probably others that have escaped the detection of this review.

Fairly high reliabilities have been reported for some of these methods (although evaluations must be tempered because calculations of reliability are not consistent).

Falloff using these systems is also relatively modest for severe events. The use of these interviews, moreover, may result in significant cost savings.

The Detroit Couples Study Life Events Method. Like the SEPRATE, this interview is a checklist life event measure, with semi-structured probes administered after an event is reported. An offshoot of this interview was utilized in the National Survey of Health and Stress (Kessler, 1990).

Several features of the LEDS were included in its development. Acute, discrete events are differentiated from chronic difficulties. In addition to standard queries about domain-specific events, questions related to event “meaning” (e.g., revelations about the character of others, disappointments, situations turning out better than expected, and upsetting news) were adopted from the LEDS (Kessler & Wethington, 1991).

A number of features were incorporated into the design of the study in order to improve recall and accurate dating of life events. The features include concrete wordings of life event questions, contextual cues (arranging event questions by domain and embedding them in sections including other sorts of questions about that domain), multiple frames of reference (e.g., questions about learning something negative about the character of another person in addition to questions about difficulties “getting along” with someone), and memory aids (a life event calendar and calendar review to correct dates).

Four category stress severity ratings were assigned, based on the “loss” and “danger” long-term contextual threat ratings from the LEDS. However, the ratings were more similar to life-change event ratings than LEDS contextual threat ratings, since the interview did not collect sufficient contextual information to reproduce the latter.

A detailed analysis of event reports in the Detroit Couples study showed that this interview measure has acceptable psychometric properties overall, with some intriguing exceptions (Kessler & Wethington, 1991). The design of the study (husbands and wives interviewed from the same couple) made it possible to calculate reliability estimates for events rated as severe. The average reliability of reports was .64 for husbands and .68 for wives. However, inspection of reliability within event types showed that events that have the potential to be embarrassing, such as criminal activity and marital separations, were less reliably reported than other events. It is difficult to imagine that such a pattern reflects a common tendency to forget about embarrassing events. A more likely interpretation is that the occurrence of these events was consciously concealed. In addition, a comparison of husband and wife reports shows that personal events were consistently reported more accurately than events that occurred to one’s spouse. The 2 to 3 percent monthly falloff rate for event reports, moreover, was lower than the 5 percent monthly falloff for a checklist reported by Funch and Marshall (1984), but greater than that reported for the LEDS (Brown & Harris, 1982).

The overall results, though, from the Detroit Couples Study are not as impressive as those obtained with the LEDS. The relative risk of onset of depression associated with severe events in the previous 12 months is 2.43 (odds ratio) compared to the range of 7 to 13 reported for the LEDS (see Paykel, 1978). It should be
noted, moreover, that the estimate is also less than that obtained by the SEPRATE.

Nevertheless, the more structured methods of the interview may have resulted in a significant cost-saving. The average interview lasted 78 minutes. Coding time for the interview, including detailed coding of the life events, was less than 2 hours. Conventionally trained survey interviewers and coders were used, and only 1 day of special training was necessary for each group. Most significantly, 1755 interviews were conducted in a 4½-month period, and the coding of all events was accomplished in 2 months, a pace that could not even be approached using the LEDS.

It is important to note that methods such as these were developed specifically for a large-scale community survey. The measures used in the National Survey of Health and Stress (Kessler, 1990), an offshoot of the Detroit Couples Study, were applied in a sample of 8,000 representative of Americans, including ethnic minorities, aged 15 to 54.

The Munich Events List. The Munich Events List (Wittchen et al., 1989) is a three-step personal interview procedure. The interview consists of 85 life event descriptions, which are concretely specified. In the first step, the interviewer and respondent read the event descriptions together, with the respondent deciding whether such an event had occurred over the recall period. In the second step, the interviewer asks questions to assess the relevant context of the events already reported. In the third step, respondents rate the impact of each incident subjectively; then the interviewer rates the events on a number of objective dimensions, including whether the event involved gain or loss, was primarily positive or negative in its impact, could have been averted by the victim, or might have been brought on by the subject's mental condition or disordered behavior. There is also an objective severity rating scale, conceptually more similar to a life change—readjustment unit measure than the severity ratings used in the LEDS. Subjective and objective ratings of severity are correlated at .67 (Wittchen et al., 1989).

Reliability estimates were calculated in a test–retest analysis, by event type and/or severity. The investigators reported nearly perfect concordance for severe events such as deaths (kappa not reported), with lower agreement for less severe incidents (kappa = .74). Falloff was also estimated to be near zero for the severest events (0.36% per month, vs. 1% per month for the LEDS). Data have not been reported with which to calculate the 12-month risk of depression or other illness onset associated with severe events.

No data are available on the relative costs involved in fielding this interview, which has been used only in Europe. Samples have been small (Wittchen et al., 1989).

The Paykel Brief Life Event List. Paykel's life event interview (1983) has been used in a number of studies. This interview consists of 63 life event and difficulty questions, and can be administered in an average of 30 to 40 minutes (Cooke, 1985). Once an event or difficulty has been reported, the interviewer uses semi-structured probes in order to obtain more objective information and to estimate its emotional impact on the subject. Events and difficulties are dated, and is determined whether they occurred before or after illness onset. They are also evaluated for their independence from the subject's illness status, whether the victim could have averted the event, respondent-reported negative and positive impact of the event, and objective contextual severity of impact, a measure adapted from the LEDS (Paykel, 1983). Interviewers are given 3 weeks of intensive training in rating events, and all events are rated by two independent raters. (It has not been reported how much the double-rating system reduces production time per interview.)

Inter-rater reliability of dimensions rated using this method appears to be adequate, ranging from .95 for specific event occurrence to .76 for objective negative impact (Paykel, 1983). In Cooke's replication (1985) inter-rater agreement on objective severity of impact (kappa = .64) is just within adequate range. The instrument has been used with adult community samples and patients.

Henderson, Byrne, and Duncan-Jones List of Recent Experiences. The LRE (Henderson et al., 1981; Steele, Henderson, & Duncan-Jones, 1980) is a personal interview method based primarily on the Tennant and Andrews (1976) life events scale. Probes administered after event responses are designed to date the event, distinguish discrete events from reports of chronic situations, obtain a brief description of the experience, and obtain ratings reflecting the respondent's judgment regarding severity of impact.

A separate staff rates events. Most of the ratings are based on the life-change weights developed by Tennant and Andrews (1976). However, for rare or unusual events for which no life-change weights are available, event weights are calculated by averaging the subjective impact reports of all sample members who reported that type of event. This averaging method is intended to reduce the confounding between subjective retrospective accounts of events and the respondent's mood.

Specific data about falloff in reporting severe events and question reliability are not available for a 12-month recall period, although 10-day test–retest reliability estimates are very good, estimated at .89 to .94 for the scale score, and .70 for specific events (Steele et al., 1980). The relative risk of onset of depression or other psychological disorder associated with severe events also cannot be calculated, since this study used a scale of "neurotic symptoms" as its outcome. However, the relative risk of experiencing a high level of neurotic symptoms among those reporting a high score on the life event scale (vs. a low level) was 2.88 (Henderson et al., 1981). The measure has been used only with adults.

Summary of Alternative Measures

Although the use of alternative interview measures of life events would substantially reduce the costs of conducting life event interviews, the present state of research on these instruments does not recommend their use without qualification. For example, the expectation that more structured life-event interviews would reduce interviewer and rating time is not necessarily fulfilled. The SEPRATE, which is one of the most highly developed of the alternative systems, requires extensive editing and expert rating upon its conclusion (Dohrenwend et al., 1993). Expert rating, as well as the preparation of event descriptions, requires a substantial investment of time that may be no less than that required to field the LEDS. The Paykel Brief Life Events List...
requires as much interviewer training as does the LEDS (Cooke, 1985); moreover, reliability estimates for the measure that maximally reduces interview costs are also disappointing (Kessler & Wethington, 1991).

Some alternative methods, however, do have their strengths. The Detroit Couples Study method, which does not require extensive interviewer training, is clearly the most compatible with existing community survey methods in the United States. The SEPRATE has several useful rating dimensions and is compatible with a widely accepted theory of the stress process. Falloff estimates for the Paykel Brief Life Events List and the Munich Life Events Inventory rival those reported for the LEDS. Unfortunately, there is no research yet published that directly compares these methods and weighs their theoretical and logistical costs and benefits.

**Shortened Versions of the LEDS**

Miller and Salter (1984) are perhaps best known for their conclusion that there is no "shortcut" for rating the severity of life events with the precision of the LEDS. Yet they also concluded that better life-event rating could be achieved through the use of more detailed standardized probing, rather than through continued reliance on unstructured probing.

In response to this statement, a number of investigators have attempted to improve the LEDS interview method by modifying its interview techniques, primarily by structuring probes for event, which would both shorten the time it takes to administer the interview and determine the event ratings. Such structuring, if demonstrated to reproduce the event ratings produced by the LEDS, would substantially reduce its costs. Three revisions of the LEDS are reviewed in the following sections.

**The Structured Life Events Inventory**

The Structured Life Events Inventory (SLI: Wethington, Kessler, & Brown, 1993) was designed as a *more structured* version of the open-ended LEDS. The purpose of creating the structured version was to experiment with ways to reproduce Brown and Harris's contextual threat ratings, while also using interviewing techniques more consistent with conventional American survey techniques (e.g., the Detroit Couples Study). Another purpose of structuring the LEDS was to reduce the amount of time necessary in order to produce contextual threat ratings, as well as to cut the amount of necessary training time.

The SLI resembles traditional survey interviews more than it does the LEDs, but it still differs significantly from them as well. Consisting of a series of questions designed to elicit events, the SLI provides instructions for the interviewers to probe some of the responses in an extemporaneous, conversational manner, that is, to follow the lead of probe questions that are suggested in the interview, but not necessarily to read the questions verbatim.

As does the LEDS, the SLI requires interviewers to make research judgments while they are interviewing. Interviewers are told that the purpose of the probing is to estimate the objective severity of an event or difficulty reported by the respondent. During the interview, interviewers make judgments about the severity of the event or difficulty the respondent is reporting, and record their summary judgment regarding severity in the interview, along with a detailed description of the situation. Because both of these tasks require fairly precise knowledge about the event or difficulty, as well as some understanding of some of the important aspects of the respondent's life that may increase the event’s severity, care has been taken to assist the interviewers in making correct judgments. Time-consuming probing of events and difficulties that are of low-moderate or less severity, as well as failing to detect potentially severe events and difficulties, are errors that can greatly reduce the quality of the data.

Although this set of interviewer tasks and judgments may initially seem very challenging, and very difficult to train and supervise, the investigators tried to reduce the burden on the interviewers in several ways—primarily by reducing reliance on memory. A training manual was produced that described the general principles used to make judgments about event and difficulty severity. The conventionally trained survey interviewers were given 3 days of training. Exercises based on examples reported in the LEDS dictionaries were used to teach the interviewers how to make judgments. Interviewers were also trained in probing techniques.

In order that training time be reduced, the interview schedule contained as many instructions as feasible to make it possible for determining severity. This was done so that interviewers would not have to memorize a great deal of specific information about what sorts of events and situations are rated as severely stressful, or less. In the interview design, two major strategies were utilized. The first principle was to begin probe sequences with questions designed to screen out as quickly as possible events that were not likely to be severely stressful. For example, reported financial difficulties were not probed if respondents had not had to deplete their savings, sell or skip payments on their major possessions, or go without medical care and necessities such as food and clothing for the family. If events were determined not to reach threshold for severity, the interviewer was instructed to skip to the next question sequence.

The second design principle was to produce for each question sequence a series of optional probes that were sensitive to contextual variations that make an event severe or not. An example of this principle is the question and probe sequence for respondent pregnancy: Did this pregnancy have any negative impact on your career or job plans for the future? (Optional: Will you be/Were you) able to keep your job? (Will you be/Were you) forced to cut back on your work hours? [IF YES] (Is/Was) this your choice or someone else’s decision? (Did you have/Will you have) to turn down a promotion, transfer to a less demanding job, or put your career plans “on hold” in any way? [IF YES] (Is/Was this your choice, or someone else’s decision?) In addition to the sequence excerpted, there are also questions about whether there were difficulties with the father or other family member during the pregnancy, financial or housing problems, near miscarriages, or other medical emergencies involving either mother or baby. Each includes a set of optional probes to gather information about potentially severely stressful circumstances surrounding the event.
In order to examine interviewer judgment and maintain consistency, all interviews are then examined and rated by an expert rater, trained in the LEDS method. Ratings for long-term contextual threat, event type, independence from respondent illness, and focus are assigned to the events, based on interviewer ratings and other information from the interview.

The SLI was fielded in an experimental study in 1992. This study was designed to examine the reliability and validity of the structured version, in comparison with the original semi-structured method. A sample of 243 community-dwelling residents were interviewed, half with the SLI and half with the original LEDS. Forty-one of the respondents were interviewed twice, using both methods. The LEDS interviewing and weekly panel consensus meetings were supervised by an expert investigator trained by Brown and Harris.

Results from this comparison study, although still under review, are promising. The average SLI interview production time (including interviewing, coding, and rating) was 9 hours, compared to 16 hours required for the LEDS. SLI interviewers can conduct more interviews per week (8) than LEDS interviewers (3–4). In addition, SLI interviewers were able to reliably distinguish events of severe long-term contextual threat from more minor events. In addition, the risk of depression onset associated with severe events and difficulties (3 months prior to onset) was similar for the LEDS and SLI in this study, with the LEDS having a slight advantage.

Nevertheless, much work still needs to be done on the SLI. Between-interviewer rating reliability was not perfect. The LEDS appears to be superior to the current SLI in eliciting and rating difficulties. In order to improve the SLI, development of a second experimental test is underway.

Costello's Two-Stage Screening Method

Costello (1982) and Costello and Devins (1988) have experimented with a two-stage screening method in order to reduce LEDS interviewing and rating time. Reasoning that some of the discrepancies found between self-administered checklist and personal interview techniques were due to discrepancies in how events were defined, the investigators developed a checklist based on the LEDS. Event and difficulty descriptions were worded to reflect contextual circumstances that increase severity ratings.

Study participants were asked to fill out a screening checklist. Within a week of completing the screening, participants were interviewed in their homes by LEDS-trained interviewers who were blind to the checklist results. The investigators then compared severe events and difficulties reported from the two methods.

The interview and checklist methods were highly consistent when a severe event or difficulty had not occurred. However, the two techniques were not as consistent when such an event had occurred; only 27 percent of those indicating a "severe" event or difficulty on the checklist were rated as having experienced a severe event or difficulty of that type in the subsequent interview. These results indicate, though, that a checklist method could be used to screen for events at the beginning of a personal interview, and that the subsequent LEDS administration could be shortened. Such shortening would save considerable interviewing and rating time.

The Edinburgh Version of the LEDS

Miller and associates (Miller et al., 1986) have also fielded a shortened version of the LEDS. As recommended by Costello (1982), this interview uses a self-administered checklist screening procedure, with the interviewer probing only those events checked on the list. A panel then assigns ratings based on the LEDS system.

In a study using this method, the relative risk of onset of psychiatric disorder (depression or anxiety) is estimated at 2.40 for those who experienced any severe event and 3.21 for those who experienced a severe event or difficulty (Surtees et al., 1986). Falloff rates for event recall appear to be very similar to the LEDS (Surtees et al., 1986). This methods appears to have shortened interview and rating time, although per-interview time estimates have not been published.

Future Directions

It is traditional to end such a review with a call for further research. There is a need for more research comparing the interview methods themselves, and more attention to documenting the costs of different personal interview methods. More critical, though, are guidelines to help researchers make judgments about which of the methods reviewed is appropriate for their projects.

Choosing the Appropriate Method

Two critical criteria are necessary for judging the appropriateness of an interview method: (1) its suitability to answer the primary research questions of the study, and (2) practical limitations.

Several unresolved issues that have circumscribed use of personal interviews based on the LEDS still remain. The first is that the rating system, although clearly defined and documented, confounds some event severity ratings with social vulnerability factors. Many have argued (e.g., Dohrenwend et al., 1993) that these dimensions should be kept strictly separate. The second issue is that the costs of administering any personal interview measure are high in comparison to a checklist. A number of researchers have therefore sought alternatives in order to keep costs down. The third issue is interviewing logistics, over and above study costs. Studies utilizing the LEDS have employed relatively small samples (usually 400 or less). Interviewers require extensive training. Post-interview rating and attendance at weekly panel meetings limit the number of interviews that can be completed per week; as a result, interviews take place only over long periods of time, or a very large number of interviewers must be hired and trained. Even if the panel method is not used to rate events and difficulties, as in the SLI (Wethington et al., 1993) and...
the Brief Life Event List (Paykel, 1983), interviewers must be closely supervised and periodically retrained. Some progress has been made in addressing the confounding issue of rating, primarily through the development of the SEPRATE. It is still unclear, though, whether the SEPRATE represents a significant cost savings over the LEDS, or reduces the logistical problems that make the LEDS less practical to apply in larger sample surveys. The ready availability of the LEDS dictionaries, the longer time in which the LEDS has been used, and its ready application to diverse populations (e.g., adolescents and other cultural groups)—all these advantages may outweigh the confounding issue of rating in some researchers' eyes.

Significant progress has also been made in reducing cost and logistical problems, primarily through the development of shortened or more standardized versions of the LEDS (Cooke, 1985). Questions still remain, however, about their replicability and reliability, since none has been used as extensively as the LEDS.

Acknowledgments
The development of this chapter has been supported by grants from the MacArthur Foundation Research Network on Successful Midlife Development to Ronald C. Kessler and USDA Hatch (321-7419) to Elaine Wethington. The authors wish to thank Sheldon Cohen, Stanislav Kasl, Bruce Dohrenwend, Bruce Link, and Martha Bruce for their comments and contributions to this chapter.

Notes
1. Training for the LEDS is conducted by Tirril O. Harris at Royal Holloway and Bedford New College, 11 Bedford Square WC1, UK; and also by Dr. Ellen Frank, University of Pittsburgh, Western Psychiatric Institute and Clinic, Bellefield Tower, 3811 O'Hara Street, 8th floor, Pittsburgh, PA, 15213, USA.
2. Training for the SEPRATE is available through a staff supervised by Dr. Bruce Dohrenwend, New York State Psychiatric Institute, Box 8, 722 West 168th Street, New York, NY 10032, USA.
3. Inquiries about use of the SLI should be directed to Dr. Elaine Wethington, Department of Human Development and Family Studies, G63 Van Rensselaer Hall, Cornell University, Ithaca, NY 14853, USA.

References


The Environmental Perspective


There are several diseases, including schizophrenia, alcoholism, and opiate addiction, for which the long-term prognosis is subject to disagreement between clinicians and researchers and also among researchers. Part of this disagreement may be attributable to a difference in the populations they sample. The clinician samples the population currently suffering from the disease (a "prevalence" or census sample), while research samples tend to more nearly represent the population ever contracting the disease (an "incidence" sample). The clinician's sample is biased toward cases of long duration, since the probability that a case will appear in a prevalence sample is proportional to its duration, hence "the clinician's illusion." The statistical mechanism of this bias is illustrated and its consequences detailed. Other sources of sampling bias in clinical and research samples are briefly described and partial remedies are suggested. (Arch Gen Psychiatry 1984;41:1178-1182)

There are several diseases for which there exists a great disparity in the prognostic expectations of practicing clinicians and investigators of the natural history of the disease. This difference in perspective sometimes leads clinicians to reject formal investigations as invalid or at best irrelevant to the patients they treat. It similarly leads some researchers to view clinicians as unable to surrender clearly invalid ideas about the nature of the problems and prognoses of their patients. Although doubtless some of these discrepancies arise from the difference in their professional roles vis-à-vis patients, another possibly major source of this difference is a function of a sampling problem that has not been generally appreciated.

In this article, illnesses in which this bias is most likely to be found are identified and a neglected methodologic problem that could account for the disparity wholly or in large measure is illustrated. In addition, several other sources of bias or discrepancy that may further complicate and obscure the problem are discussed. Finally, this same problem is shown to be reflected in inconsistencies among research findings. Current research reports in psychiatry often do not allow a proper consideration and correction of the bias. Recommendations for future research and research reports are made.

A 1982 example of a disputed prognosis is available for alcoholism. Although it is not the intent of this article to draw substantive conclusions, it seems likely that the debate regarding the long-term outcome of treatment programs for alcoholics has been fueled by the conviction of many clinicians that the research findings of the studies in question did not coincide with their clinical experience. Nevertheless, the preponderance of the research findings, with regard to the long-term outcome of alcoholics seems to support the Sobells' findings in a general way. That is, they indicate that at least half of all serious alcohol abusers recover completely, many without formal treatment. Furthermore, in contrast to conventional clinical wisdom, a fair proportion of these once-alcoholics have returned to social drinking rather than to abstinence.

A similar debate exists with regard to prognosis in schizophrenia. The traditional view is that schizophrenia is a chronic, deteriorating, and more or less incurable illness. Among clinicians, the most favorable outcome estimate is the "1/3, 1/3, 1/3" rule: one third will get better, one third will stay the same, and one third will get worse. The DSM-III of the American Psychiatric Association is even less sanguine, stating that "a complete return to premorbid functioning is unusual—so rare, in fact, that some clinicians would question the diagnosis." Thus, the issue of prognosis is addressed, in part, by defining the illness in such a way as to resolve it. In sharp contrast, studies of the natural history of schizophrenic illness suggest a much more favorable prognosis than the "one-third will get better" rule.

There are a number of other conditions in which a parallel discrepancy exists in the prevailing view of clinicians and results from clinical trials when compared with the evidence produced by research into the natural history of the condition. For example, Schacter has found what many lay people know firsthand, namely, that many people are obese at some time in their life but return, with or without professional help, to a weight within or close to published normal standards on a more or less permanent basis. Similarly, in contrast to the experience of services designed to help people quit smoking, it has been found that more than 60% of once heavy smokers who tried to quit no longer smoked.
Perhaps even more striking is the view of clinicians, widely shared by the public, of opiate addiction as an intractable state for most, if not all, users. This view was forcefully contradicted by Robins and associates, who found that of a sample of Vietnam veterans who were addicted to heroin when interviewed after their return to the United States, 73% were drug free 2½ years later, often without great effort. Of all those who became addicted in Vietnam, even a larger proportion, 88%, avoided relapse over the three years following their return.

It is not our intent to come to any substantive conclusions about this article, however, and readers who disagree with our characterization of findings in any given area are requested to consider the methodological points being made and their relevance to research in areas familiar to them. Disparities similar to those discussed previously may be expected in investigations of recidivism of juvenile delinquents and adult criminals as well. They may also be found, although probably to a lesser degree, in certain more strictly physical illnesses, such as chronic pain conditions, arthritis, and some cancers, and possibly in many other diseases.

Are clinicians such poor observers? Of course not. They are simply reporting a reasonably accurate abstraction from their experience in the treatment of these conditions. Their day-to-day experience consists of encounters with a group of patients who seek their help. Within this experience are two sources of bias, the census nature of the group and the fact that these cases are necessarily drawn from that subset of patients who have sought help from a particular service provider. The biases attributable to the help-seeking behavior will be discussed later in this article and have been alluded to or discussed in many other places.

However, the bias due to the census nature of the sample is likely to be less familiar and of even greater consequence in a number of settings, and is therefore the major component of what we call the “clinician’s illusion.”

The diseases most likely to be subject to the phenomenon of large discrepancies between clinicians’ views of the disease and researchers’ findings are characterized by great variability in the duration of the illness. Other characteristics (which serve to make long duration possible) are that they are not immediately life threatening and that treatment is either confined to symptom suppression or, if curative in intent, is often unsuccessful.

To illustrate how the variation in duration of the illness may affect the clinician’s view, look at a hypothetical disease for which the treatment provides only symptomatic relief. To make the example even simpler, without any loss of generality, let us assume that all persons will be treated for the entire duration of the illness. Suppose further that the distribution of duration of illness in the population of all persons who ever contract it is as displayed in Fig 1.

As can be seen, the distribution posed in Fig 1 is skewed, with a median duration for the entire population of less than six months and one tenth of all cases lasting more than two years. Although this is likely to be a reasonable distribution shape for some of the illnesses discussed, a more symmetrical distribution may be characteristic of others. However, the clinician’s illusion is not dependent on this shape but only on the degree of the variability of duration relative to the average duration (literally, on the coefficient of relative variation).

Imagining a clinician viewing persons with one of these illnesses as they appear in the current case load, say, on ward rounds on a given day. The probability of encountering a given patient on this day will be proportional to the duration of this patient’s illness. A patient whose illness lasts a year is far more likely to appear on any given day than one whose illness lasts only two months, in fact, six times more likely; and a patient whose illness lasts for ten years is ten times more likely to appear than one whose illness lasts only one year. For this example, each of the few patients making up the 2% of the population of longest duration (X = 128) is 64 times more likely to appear than each of the 40% of shortest duration (X = 2), 32 times more likely to appear than each of the 25% of next shortest duration (X = 4), and so on.

The outcome of this duration-dependent differential probability of being encountered on rounds is that the distribution of illness duration for the rounds sample is markedly different from that of the population made up of people who ever contract the illness. By weighting the percentages in each duration interval in Fig 1 by its mean and recomputing the percentages for the weighted distribution, the duration distribution for the rounds sample of Fig 2 is produced. Perhaps most startling is the outcome of this differential weighting for the typical current clinic group; nearly a quarter of the group is made up of the 2% of the population who have the longest duration, i.e., the worst

---

**Fig 1.**—Distribution of illness duration in population. (Sum of X times percent values for seven intervals in Fig 1 is 1,068. This is now divisor for weighted percentages. For example, percent of lowest interval for rounds sample is 2[40]/1,068 = 7.5%; for next highest, 4[25]/1,068 = 9.4%; etc.)

**Fig 2.**—Distribution of illness duration in clinic sample from same population.
prognosis. Furthermore, the ultimate average duration for the daily clinic rounds sample seen by our hypothetical clinician will be more than four years (ie, 50.8 months), in stark contrast to the true population average duration of less than one year (10.7 months). Small wonder that the clinician will view any truly population-based findings as grossly discrepant from daily experience. Note again that the phenomenon does not depend on the shape of the distribution but only on its relative variability.

It is useful to examine what happens to this illusion if a duration requirement is inserted into the diagnostic criteria, as has been done for some of the diseases discussed previously. Suppose in our previously mentioned example that those ill less than six months were excluded from the diagnosis. For the remaining 35% of the entire population, the mean duration is now 25.4 months. A clinic sample of cases so diagnosed would have a mean duration of 60.5 months, still a large discrepancy.

How then can one correct the observed rounds sample for this distortion of its population representation? Since each case's chance of being included in the sample is proportional to its illness duration, one needs to weight each case inversely by its duration to reproduce the population distribution. Although we do not know what the eventual duration of each case will be, we do know the duration up to the time of observation, and Freeman and Hutchison provide a detailed exposition of how to use durations to date of the clinic sample to estimate the distribution of illness duration for the entire population.

In the previously mentioned example, we assumed a constant probability of treatment throughout the illness. Should this not be the case, the bias may be increased or decreased. If treatment is most likely for the short duration cases, the bias will be lessened, if most likely for the long duration, it will increase. Probably the latter is most frequent in real clinical populations. Again, it is a question of the ratio of probabilities. In our example, if the longest duration group were twice as likely to be treated relative to the shortest, they each would be 128 (rather than 64) times more likely to appear in the sample than a member of the shortest duration group.

One necessary consequence of this bias in the representation in the sample is a bias with regard to the presence of any condition or symptom more characteristic of long-term than short-term treatment of patients. Therefore, it may be impossible to distinguish between a characteristic affecting the duration of illness (failure to recover from a causally related to the illness).

In sum, the clinician's illness is the attribution of the characteristics and course of those patients who are currently ill to the entire population contracting the illness. To put the issue in terms familiar to epidemiologists, it is the consequence of using a prevalence sample as a substitute for an incidence sample.

OTHER SOURCES OF BIAS IN SAMPLES FROM CLINICAL SETTINGS

There are, of course, several other reasons why samples of cases in treatment settings are likely not to be representative of the entire population of persons experiencing the condition in question. Although they have been discussed in other places, it is useful to consider them herein because they are likely to add to the illusion identified previously.

First, those who recover from a given episode of the condition without entering the formal treatment system, that is, by using personal and social resources, are likely to have a better prognosis by virtue of the stability of the curative factors. Unlike those who seek professional help, their help will not disappear when they recover, but will tend to remain through the improved state. A robust constitution is more reliable than the pharmacopeia.

Those who enter the formal system are likely to have been ill longer. For some of the conditions we have discussed herein, the patient's "social margin" will have declined over time and be accompanied by a decline in the natural support system and increase in alienation from and rejection by others. For these and other reasons, past duration of illness may be causally related to duration of illness.

The clinician's illusion is also related to another common observation, at least in the field of schizophrenia, and probably in some or all of the other diseases mentioned. This is the apparent association between disease onset at a young age and poor prognosis. To illustrate how this observation could be an artifact closely linked to the previously identified bias, let us examine the characteristics of any given age group seen by our hypothetical clinician, say, all males, aged 20-25 years. Note that those who became ill at age 20 years or before and have already had a longer illness duration than those with later onsets, and therefore can be expected, on the average, to have a duration that projects longer into the future. Or to look at the same phenomenon from another angle, we can note that those who recovered during the decade following their onset at age 20 years (the patients with early onset and good prognosis) are no longer present to be observed. Therefore, if there is any diagnostic value to duration of past illness as such, this (remaining) group is, on the average, a group with more severe illness.

Note that this is in the absence of any true biological or any other implication of age of onset for the disease process. However, one reason, let us insist that we are not trying to present or evaluate empirical evidence on any issue but only to illustrate problems that arise with empirical findings that are inaccurately interpreted.

To return to other reasons why a clinical population might be atypical compared with the entire population, we note another biasing artifact known as Berkson's fallacy. It is generally the case that those who have other disabilities that are not causally connected to the condition being investigated are more likely to enter the formal treatment system. Thus, a clinical group may show a conjunctive of problems or conditions that are not uncorrelated in the population. To use Berkson's example, those whose illnesses are more likely to be found in the hospital than those with diabetes alone. Similarly, patients who are depressed and also abuse alcohol may be more likely to enter into treatment than those with either affliction alone. This second error source is a kind of ascertainment bias that, like the others mentioned herein, will further confuse the picture and confound sound conclusions.

This problem should not be confused with that of the distortion of relationships between characteristics that is typical when samples are selected on the basis of these characteristics. When, for example, two characteristics are positively correlated in the population and selection into a sample is based only on the presence of either (or both) of these characteristics, their correlation in the sample will be negative. If other, unrelated characteristics are also involved in selection, the sign of the correlation may not change, but its magnitude will be diminished. Thus, selection for symptoms is yet another source of discrepancy between clinical samples and the populations whence they come.

FACTORS RELATING TO THE MAGNITUDE OF THE CLINICIAN'S ILLUSION

In addition to characteristics of patients and treatment selection that lead to bias in treated samples, it is worthwhile to identify factors that are likely to maximize the impact of the clinician illusion. The disparity in the clinician's and the community observations and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

THE SOURCE OF THE ILLUSION

Distortion of illness in clinical populations is by tinations, there is a point of full re.

5. Treatme of the clinician's illusion is also related to another common observation, at least in the field of schizophrenia, and probably in some or all of the other diseases mentioned. This is the apparent association between disease onset at a young age and poor prognosis. To illustrate how this observation could be an artifact closely linked to the previously identified bias, let us examine the characteristics of any given age group seen by our hypothetical clinician, say, all males, aged 20-25 years. Note that those who became ill at age 20 years or before and have already had a longer illness duration than those with later onsets, and therefore can be expected, on the average, to have a duration that projects longer into the future. Or to look at the same phenomenon from another angle, we can note that those who recovered during the decade following their onset at age 20 years (the patients with early onset and good prognosis) are no longer present to be observed. Therefore, if there is any diagnostic value to duration of past illness as such, this (remaining) group is, on the average, a group with more severe illness.

Note that this is in the absence of any true biological or any other implication of age of onset for the disease process. However, one reason, let us insist that we are not trying to present or evaluate empirical evidence on any issue but only to illustrate problems that arise with empirical findings that are inaccurately interpreted.

To return to other reasons why a clinical population might be atypical compared with the entire population, we note another biasing artifact known as Berkson's fallacy. It is generally the case that those who have other disabilities that are not causally connected to the condition being investigated are more likely to enter the formal treatment system. Thus, a clinical group may show a conjunctive of problems or conditions that are not uncorrelated in the population. To use Berkson's example, those whose illnesses are more likely to be found in the hospital than those with diabetes alone. Similarly, patients who are depressed and also abuse alcohol may be more likely to enter into treatment than those with either affliction alone. This second error source is a kind of ascertainment bias that, like the others mentioned herein, will further confuse the picture and confound sound conclusions.

This problem should not be confused with that of the distortion of relationships between characteristics that is typical when samples are selected on the basis of these characteristics. When, for example, two characteristics are positively correlated in the population and selection into a sample is based only on the presence of either (or both) of these characteristics, their correlation in the sample will be negative. If other, unrelated characteristics are also involved in selection, the sign of the correlation may not change, but its magnitude will be diminished. Thus, selection for symptoms is yet another source of discrepancy between clinical samples and the populations whence they come.

FACTORS RELATING TO THE MAGNITUDE OF THE CLINICIAN'S ILLUSION

In addition to characteristics of patients and treatment selection that lead to bias in treated samples, it is worthwhile to identify factors that are likely to maximize the impact of the clinician illusion. The disparity in the clinician's and the community observations and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:

1. The variance in illness duration is large relative to the mean duration.

2. Disease manifestations fluctuate, symptoms are episodic, and treatment is sporadic and handled by multiple clinical services or clinics. Under the assumption that the clinician or clinician's observers and the true state of the population will be greatest when the following occur:
3. The history of the illness, its onset, and prior treatment are secured by unclear definition ("insidious onset") or by failure of the system to attend to them.

4. Treatments provided by the clinician reduce the intensity of symptoms rather than effect a complete cure. Under these conditions, there is less need or motivation to follow the patient to the point of full recovery.

5. Treatment is provided by specialists who are unlikely to counter the recovered patient in the context of different problems or routine care.

THE CLINICIAN’S ILLUSION AND DISPARITIES AMONG RESEARCH FINDINGS

Distortion of population characteristics by the unequal occurrence in currently ill samples of patients with different duration of illness is by no means a phenomenon limited to clinicians. It has provided a familiar convenience to accept persons with some established diagnosis who attend some clinical service as representing a larger population of persons with the disorder being investigated. It may be recognized that patient selection may be affected by the service system, that the patients may differ from those in other geographic regions, and that in any case, they are not a randomly sampled, definable population. Nevertheless, these sources of selection bias are not often considered as seriously invalidating the conclusions drawn from the investigation. This expectation of validity may be often or even usually justified. That is, when the conclusions are suitably limited and appropriate for those patients who are being seen in services like the source of the sample, and no more general inferences about the disease as such are drawn, such studies are useful.

However, Fletcher and Fletcher's review of 30 years of medical research articles suggested that data necessary to describe the sample, and therefore the population to which generalization may be appropriate, are increasingly absent from published reports of research on medical conditions. In spite of this inconsistency in reporting relevant sample descriptors, researchers are often concerned about the great variability in outcome observed in these populations. Indeed, these concerns have led to increased efforts to improve the diagnostic system for mental illness. The frequent admonition that inconsistencies in outcome cannot be expected to disappear without valid, consistent, and rigorously applied diagnoses should be accompanied by an equal concern for population definition.

Not only are samples drawn from well-defined current cases likely to be heterogeneous and biased with regard to duration of illness, but even samples of new inpatients are treated in a similar fashion. For example, it has been found that patients diagnosed in the same hospital as first admission to a psychiatric facility had manifested psychotic symptoms for durations varying from less than one week to more than five years. Furthermore, duration was substantially correlated with the nature of the symptoms.

Another by-product of the clinician's illusion is an overestimate of the number of new patients entering a treatment system. It is, of course, possible to estimate incidence from prevalence figures, given the validity of certain assumptions. In the absence of explicit estimates, those operations clinical facilities for chronic diseases are likely to overestimate the number of new cases. This problem, which causes an "accrual shortfall" in studies based on new cases, is apparently endemic to research on chronic or recurring disease processes.

It is not necessary to invoke a treatment setting in order to find the kind of biased representation we have discussed. As must follow from the fact that in our illustration we assumed that all those with an illness were being treated as long as they were ill, the phenomenon will be present in untreated or "true" prevalence studies as well. These studies will, of course, avoid the problems associated with treatment settings, as enumerated previously. However, they will still overrepresent the long-term ill. This problem may not seem so serious because it is often thought that an understanding of the more lasting forms of an illness is the most urgent need. However, in comparing such a biased sample with controls, there is likely to be no way to separate characteristics that have an effect on the duration or course of an illness from those with a causal connection per se.

This problem of biased sampling needs to be kept distinct from the more familiar problem for preferring incidence to prevalence samples, namely, the difficulty in the latter samples in distinguishing causes of the disease from consequences. Unfortunately, incidence samples of relatively rare diseases are difficult to locate; longitudinal samples of populations must be impractically large to include enough cases to examine. For this reason, for example, the "high-risk" designs for following children at genetic or other risk of contracting a disease are employed. Even here, in theory, one must beware lest parents are selected not only for the illness but also for traits that are etiologically irrelevant but affect the course of the illness, such as intelligence or criminality, which they may pass on to their children.

One real possibility for many of the diseases discussed previously is that these etiologically uncorrelated traits may have a pernicious influence on the course of the illness. Thus, low intelligence may not affect one's risk of becoming a felon but, once a felon, may adversely affect the likelihood of successfully changing vocations. Similarly, inadequate social skills and friendship bonds may have no causal connection with schizophrenia but may impair opportunities for reality testing and improving adaptive skills in those in whom thought disorder develops.

Research findings will be most distorted when based on a point prevalence of treated patients, as opposed to a prevalence of untreated patients, one including all identified cases over a span of time. A sense of how extremely different these populations may be is suggested by the following illustration.

At any given moment, about 15,000 New York City patients are being treated in facilities sponsored or licensed by the New York State Office of Mental Health. Furthermore, there are about 15,000 annual admissions to these facilities. If one examines the census of these treatment facilities, that is, takes a point prevalence of treated patients, about 95% of the patients will have been hospitalized for more than one year (the average length is more than 20 years). At least 60% of these patients will still be there if another census is examined two years later. On the other hand, if one examines the period prevalence, approximately 30,000 patients treated during the year (more exactly, 95% of 15,000 = 14,250 plus 15,000 admissions = 29,250), one will find a very odd-shaped distribution with large peaks in the three-weeks-or-less and 20-years-or-more treatment duration groups, perhaps a third of the sample in each of these categories and the other third strung out in between.

If one examines intakes alone, the median duration may be about three weeks and only a tiny proportion will be in treatment two years later. In other words, these two treatment groups, the census on a given date and the intakes over a year's time, have only about a 5% overlap. These figures are, of course, approximate and depend on the nature of the treatment system.

If all treated patients in one year were included regardless of treatment provider, these figures would change; they also will vary depending on the mix of outpatient and inpatient service. The moral of the story is, however, independent of these details. It is that without a careful specification of the treatment history and the nature of selection, the population to which research findings can be generalized is unknown. Current conventional standards for research reports are quite inadequate for this purpose.

COMMENT

The phenomenon described herein as the clinician's illusion should be understood in the context of other illusions; that is, as a natural consequence of a combination of certain human perspectives and information-processing tendencies. Thus, it is not intended to be a pejorative term and does not connote delusion; nor is it asserted that the basic problem is entirely unfamiliar to workers and methodologists in clinical fields. Nevertheless, the possible magnitude of the effects on prognostic inferences may well have been insufficiently appreciated in the published literature. Therefore, the following recommendations are offered.

First, researchers should be careful to report the duration of illness and/or number of prior episodes in studies of...
clinical populations. The advisability of studying new cases rather than all current cases should be given careful consideration (but note the risk of accrual shortfall). When the course of the illness is at issue, the prior illness duration must be taken into account. If the past duration is long relative to the follow-up period for any of the patients, an appropriate consideration of the possible effects of this variation on the study conclusions is clearly required.

Second, natural history investigations of cases unselected for duration of illness (and preferably also regardless of treatment) should be considered crucial to the understanding of any illness whose duration varies materially relative to its average length. Such studies are demanding of continuity of effort and support over an extended period of time.

Finally, given the likely distortion of our current knowledge base, we should be cautious in advising lay people in matters relevant to the natural history of chronic diseases. This issue arises with particular potency in the area of genetic counseling. To illustrate, in a research carried out in one of the most sophisticated and research-oriented centers in the country and reported in a journal devoted to research in this area, a parent of a schizophrenic patient is quoted as saying "I've known people who were cured of cancer, of all these illnesses, but I know of no one who was cured of schizophrenia." One would hope that someone concerned with the care of this patient had informed the parent of the high spontaneous remission rate generally found over a more protracted period of time. Nor was any lifetime perspective apparent in the report of the genetic counseling offered. Wider appreciation of the phenomenon we call the clinician's illusion should prove salutary for both clinical practice and the interpretation of research findings.

This investigation was supported in part by grant MH 36971 from the National Institute of Mental Health.

Albert Stunkard, MD, gave encouragement toward the completion of this article.

References