Measuring Daily Events and Experiences: Decisions for the Researcher

Arthur A. Stone SUNY–Stony Brook
Ronald C. Kessler University of Michigan
Jennifer A. Haythornthwaite National Institute on Aging

ABSTRACT There has been a burgeoning interest in studying daily events and experiences. This article discusses a variety of methodologic challenges that face daily event and experience researchers. The issues discussed include techniques for measuring events, the development of event checklists, sampling event content, specifying event appraisals, event validation procedures, and the creation of summary measures derived from event checklists. Procedural issues discussed include determining the number of observations and persons needed for daily event studies, the evaluation of response, attrition, and missing item bias, and problems linking event reports over time.

As both the empirical and commentary articles in this special issue attest, the use of daily event methodologies in the study of behavioral phenomena has yielded compelling results. Findings from these studies completed during the last decade have expanded our understanding of the impact of the psychosocial environment and, by their typically prospective designs, have facilitated causal interpretation of microprocesses underlying daily experiences.

The authors appreciate the support of the following grants in the preparation of this manuscript: Grant R01-MH39234 awarded to Arthur A. Stone; and MERIT Award M01-MH42714, Research Scientist Development Award K01-MH00507, and Grant R01-MH41135 awarded to Ronald C. Kessler. All grants are from the National Institute of Mental Health. Requests for reprints may be addressed to Arthur A. Stone, Department of Psychiatry, SUNY–Stony Brook, Stony Brook, NY 11794-8790, or to Ronald C. Kessler, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106-1248.

Journal of Personality 59:3, September 1991. CCC 0022-3506/91/$1.50

In addition to generating new interest in the role of events and perceptions of the microenvironment, the study of daily events has presented researchers with a host of challenging methodologic issues. We first discuss two general methodologies that have been used to record daily events and experiences, the experience sampling method and the event checklist method. Variations of each method are discussed, including open- versus close-ended techniques, telephone versus paper-and-pencil recording modes, and computer-aided techniques. We then describe the development of event checklists, including issues pertaining to event sampling, level of event specificity, event appraisals, event validation and reliability, and summary measures derived from checklists. Methodologic and procedural considerations for conducting daily event studies are discussed next. We show how the number of observations and participants, the duration of time for recording events, and participant recruitment and retention procedures can affect the success of a design. We then discuss the problems we encountered when conducting these studies, including nonresponse and attrition bias, response decay, missing item bias, and the problems in linking stress reports over time. Finally, we discuss directions for future research in daily events and experience.

Our purpose in this article is to provide guidelines for conducting diary studies of daily events. These guidelines should not be interpreted as hard and fast rules. Rather, they represent our thinking about the issues and options for handling the issues. Throughout the article we also comment on the methods employed in other articles included in this special issue.

There are many levels of measurement available for characterizing experience and events. Minute-to-minute reports are obtained with experience sampling methodologies, while at the opposite extreme, events that transpire over long time periods can be measured with daily, weekly, or monthly event checklists. The duration of the event or experience measured (*event duration*) should not be confused with the period of time over which participants report events (*recording period*). It is possible to ask participants to record experiences (duration measured in minutes) for the entire day at the close of a day (recording period measured in hours or days) or to report life events (duration ranging from hours to days) for the last year (recording period measured in months). The relationship between the event duration and the recording period is an important methodologic consideration that will

be referred to later in this article. For the sake of simplicity, we will use the term "event" throughout the article to refer both to experiences (the term used by experience samplers) and to events (the term used by checklist investigators).

Methods for Recording Events

Strategies for developing event recording methods should be tied to the questions being investigated. One broad class of questions involves the amount of daily event stress experienced by participants. These questions demand use of a methodology that reasonably samples all of a day's stressful events. Another class of questions involves participants' reactions to particular events that occur during the day. One method is to record a single or a few events that meet some criteria set forth by the investigator. Events that are upsetting, pleasant, out of control, or in a particular content area (e.g., family, work) have all been used for selection. In this case, the totality of the day's experience is not represented, but rather a detailed description of event(s) relating to specific hypotheses is explored.

Two related methods of recording daily events will be discussed. The first, the Experience Sampling Method (ESM), employs a very brief recording period and a method for studying individuals in their natural environments. Participants are typically "beeped" by an electronic device (e.g., a computer strapped to the wrist or belt) at random or prespecified times during their waking hours over the course of several days. When beeped, they complete a questionnaire designed to describe their immediate experiences. Studies have measured a broad range of experiences, and reviews of the application of this method have argued for its reliability and validity (Csikszentmihalyi & Larson, 1987). An excellent review of the ESM and the details of its application is provided by Hormuth (1986).

The second method is recording events, often with checklists, at some regular interval. Although the recording period varies with the methodology, it is typical for participants to record events once a day, once a week, or even once a month. Many kinds of methods are available for recording events, and the following sections discuss characteristics of the methodologies employed for both the ESM and checklists, although, for simplicity, issues are often discussed in terms of checklists.

Open-ended methods. The mechanics of recording daily events are diverse, each having strengths and weaknesses. A straightforward method is to have participants record the events of the day using an open-ended response format. A participant may describe the event in his or her own words in several blank lines provided in the questionnaire. Participants record a certain number of events in open-ended fashion or record the event that was "highest" on some characteristic, for instance, the event that was most stressful during a day. Emmons (1991) employed a combination of these two approaches (two events which "most influenced mood") as did Campbell, Chew, and Scratchley (1991). An even simpler method is to ask whether or not any event (usually undesirable) occurred during the day, as Eckenrode (1984) and Verbrugge (1980) have done in their diary studies.

Advantages of the open-ended method are the ease and brevity of completion for participants. It is an especially attractive strategy in surveys that use telephone data collection, since it creates a conversational style of interviewing that most respondents find more enjoyable than fully structured interviews. An important consideration in the use of open-ended methods is that participants have no prompts to jog their memories of the day's events. Recall of events is probably considerably more difficult and prone to bias than recognition. This comment applies less to major events, which are not likely to be forgotten, than to minor, but potentially important events. Another consideration is the variation in the detail of event descriptions that will be observed among participants. Some participants will be telegraphic in their descriptions; others will provide much detail. Short descriptions may make it difficult for the researcher to understand the content and importance of the event, and may impede its subsequent classification.

Open-ended responses can be coded into a comprehensive, structured checklist after the interview has been completed, without burdening the respondent with the task of having to go through the checklist. This strategy relies heavily on having comprehensive responses to open-ended questions about the precise nature of each event, which means that it is likely to be infeasible in a paper-and-pencil data collection. Nonetheless, the method is used and is often effective in the development of structured questionnaires (see below).

The open-ended methodology also leaves the selection of the events under the control of participants, and increases the likelihood that personality could interact with event selection to bias the event data. If, for

instance, two events meet a selection criterion (e.g., "stressful") and one is more psychologically noxious to the participant, he or she may report only the less noxious event or perhaps the more noxious event. Campbell and associates (1991) minimize the influence of self-esteem on participants' descriptions of events by training them to provide objective event descriptions.

Checklist methods. The most popular method of assessing daily events is by event checklist. A study of negative events in a particular role domain may include a checklist of very specific events that occur frequently in that domain. A study focused on a particular outcome, like depressed mood, may include a checklist of daily events known to be important for that outcome. The level and basis of aggregation would be influenced by the focus. Holmes and Rahe (1967) provided the first model of an event checklist in the form of the Schedule of Recent Events (SRE). Checklist assessments of daily events evolved from the SRE in a natural way. They include MacPhillamy and Lewinsohn's (1975) Pleasant Events Schedule, its counterpart, the Unpleasant Events Schedule (Lewinsohn, 1975), Kanner, Coyne, Schaefer, and Lazarus's (1981) Hassles Scale, and Zautra, Finch, Reich, and Guarnaccia's (1991) Inventory of Small Events.

Mode of data collection. The majority of event checklist data is collected by self-administered paper-and-pencil data entry, although it is possible to administer telephone interviews to obtain the same data. There are benefits and drawbacks to each approach. Paper-and-pencil administration has two advantages over phone administration: It is less expensive and it is more flexible. The inflexibility of phone administration leads to problems in the sampling frame, e.g., exclusion of respondents who do not have a phone (about 8% of the total U.S. population, according to recent figures) and exclusion of respondents on days when they are difficult to reach by phone (such as during vacations or while away on business trips). It also makes it impossible to do experience sampling.

The advantages of paper-and-pencil administration are best achieved through careful respondent training and thorough monitoring of mail returns and supplemental phone recontact. Daily mailing of completed diaries is important here, and respondents should be contacted by phone whenever their diary is not received. Respondents who report that the diary is not in the mail are interviewed by phone. Phone contacts can

also be made to obtain information about missing information in a completed diary, including more detail about open-ended responses.

Daily phone interviews, on the other hand, have higher response rates than daily diaries. Data are recorded more completely in phone interviews than in self-administered diaries because the interviewer makes sure no questions are skipped. The researcher also has more control over the context of data recording in phone interviews (e.g., whether the respondent is paying full attention to the diary completion task, whether diaries are completed the same time every day, etc.). Telephone interviews can also influence the quality of open-ended responses by probing incomplete or unclear responses. Finally, phone administration provides rapid feedback to the researcher about nonresponse (i.e., a missed phone appointment or a refusal to complete the phone interview), which makes it possible to implement special efforts to complete the interview (e.g., extra callbacks to contact a participant who missed an appointment, special refusal conversion procedures). None of the diary studies in this special issue employed telephone collection of event data, although several maintained telephone contact with participants during the recording period.

Another mode of data collection uses computerized diaries, which are especially well-suited to ESM research, because computers can be programmed to signal participants at various intervals for recording experiences (for instance, hand-held computers have successfully been used; Paty, Shiffman, & Kassel, in press). The method has several advantages over paper-and-pencil methods. More information can be obtained in a comparable recording time, since the computer program guides the participants' responding. Responding trees can begin with yes/no questions and continue with more detailed questioning when an affirmative response is provided or branch to the next response tree when a negative response is provided. This allows for complex information to be recorded, which is not possible with paper-and-pencil recording methods. (Telephone interviews can also employ branching strategies by providing the interviewer with detailed, branching protocols.) Data collection is automated, which saves data entry time. Since most computers have internal clocks, time of responding can be automatically entered and length of response time noted. Rather than daily mailing of questionnaires, telephone modems can be used to transfer information. Alternatively, regular visits by the investigator to the participant, or vice versa, can be used to monitor compliance directly.

The daily events researcher should carefully consider the advantages and disadvantages of the telephone versus written versus computerized assessments. Sometimes there will be a clear conceptual reason for choosing one method. Other times the choice will be directed by fiscal limitations. In any case, the strengths and weaknesses of the assessments must be weighed in light of a study's research questions.

Development of Event Recording Methods

The issues discussed in this section concern the development of event checklists, but also pertain to the checklists used with ESM assessments and to the categorizations of open-ended responses after they have been recorded.

Sampling events and experiences. An immediate problem for the developer of an event checklist is to decide how the list will ultimately look: How many items will appear on the list? How will they be selected? What level of event specificity will be used? Will participants record events not mentioned on the list, and if so, how? How does the researcher know he/she has properly sampled events for the checklist? Should the sampled events be objective? Some of these questions have relatively straightforward answers based on logical considerations; other questions have no clear-cut answers.

Concerning how many items are on a list, the amount of time a respondent can reasonably spend completing the checklist may dictate the maximum length. In studies where events are assessed over many days, a short checklist may be in order. A list requiring 45 minutes to complete is likely to result in excessive missing data and high attrition rates. With a well-organized format, participants in a daily study were not overburdened by an 80-item checklist that could be completed in under 10 minutes (Stone & Neale, 1982). Longer event checklists are feasible if they are to be completed less frequently. Event checklists with hundreds of items can be completed once a week without distressing participants, although checklists of this length usually are not used in longitudinal studies, even when the recording period is relatively long. Again, the major consideration here is task difficulty.

Even the most carefully constructed checklist cannot include every event imaginable. Hundreds, perhaps even thousands, of events characterizing daily experience could be generated, and this would simply be

too unwieldy a list for most applications. The result is that some events related to the outcomes of interest will be missing from the checklist, and prediction of outcome from events will be less than optimal when some participants experience events not included in the checklist.

The event incidence rate that is considered high enough to demand inclusion in a checklist can also be an issue. Certainly, the event "argument with spouse" has a place in a daily event checklist since it occurs with a fairly high frequency in marital samples. But what about the event "automobile accident?" This is probably a significant event for most people, yet its very low incidence may argue for not including it in a checklist. If, for example, the checklist was characterizing daily stressfulness, then the investigator might reasonably be concerned that considerable error variance will be added to the study if the event is omitted. However, it will not be known when people do have automobile accidents, and any psychological or physical effects of the accident will not be properly attributed to the event. Some event researchers have included only commonly experienced events in their checklists (e.g., Bolger & Schilling, 1991), often to limit the recording task to reasonable lengths. Stronger associations would almost certainly have emerged in these studies if a more complete checklist had been used.

Even with a long checklist, a researcher can always imagine many events that are not on a checklist. This raises the question of how to allow participants to report nonchecklist events. An obvious and often-used solution is to include an "other" category at the end of the checklist. This technique has been especially important for the short event checklists. The problems with this combined checklist and open-ended format are the same as those for the open-ended methods mentioned above. Nonetheless, the open-ended format may provide important feedback to the researcher about omitted events that could be useful in subsequent revisions of the list.

There is another "hybrid" approach to the length versus content problem that has not, to our knowledge, been used in the event assessment area. General content areas of events could be listed in a questionnaire, perhaps with examples, to prompt participants' memories for events occurring within the content domains. These events could then be written in under the headings. Probably under a dozen content headings could capture most of the content seen in daily event checklists. The major advantage of this approach is that the participant does not face a list of hundreds of events, yet is not restricted to responding to a

limited number of specific events. It is not clear whether this approach would work in the life event area because the memory prompts may be too general to effectively aid memory. But the approach is worth researching because it could provide an assessment that is brief and yields accurate event reports.

Checklists developed for the ESM usually do not attempt to characterize all events that occur on a day. Based on his own work on relocation, Hormuth (1986) developed a checklist of situations that included 13 physical locations, 12 possible social interactions, 22 activities, and 19 topics of conversation. Diener and his colleagues (Diener & Larsen, 1984; Diener, Larsen, & Emmons, 1984) had participants categorize situations on a social dimension, a work/recreation dimension, and provide a novelty rating.

How does an investigator know that a checklist samples the appropriate domain of events for the population to be studied? This question concerns the domain of event sampling. While the SRE content was "armchaired," investigators today are considerably more demanding regarding event domain. A straightforward approach is to sample events from the participant population in which the checklist will be used. This approach has been used in the development of a daily event checklist (ADE; Stone & Neale, 1982). Using an open-ended format, participants recorded events that either "evoked an emotional response" or were "meaningful." These criterion concepts were chosen to obtain a wide range for events that might be related to the concept of stressfulness. Several thousand daily events were sampled from community members (the group where the checklist was to be used). Several slots were available for writing these events opposite different slices of the day (three slots for the slice 7 A.M. to 9 A.M., for example). This procedure assured that the ADE had representative event content. It is not known, however, whether the kind of thorough sampling of events described for the ADE is necessary to achieve an acceptable level of representativeness. A much smaller number of participants from the targeted population studied briefly might produce an adequate sampling of event content. And certainly the method suggested earlier, wherein broad content headings are employed to prompt recollection of events, does not require extensive event sampling. A comparison of events elicited by an open-ended "other events not included above" section of a checklist with the checklist events could provide information about the thoroughness of event sampling. Those events recorded with substantial

frequency in the open-ended section are candidates for inclusion in a revised checklist.

Another issue concerning event sampling is that of the objectivity of the events in a checklist, an issue that has been discussed by prominent researchers (B. S. Dohrenwend & B. P. Dohrenwend, 1974). It is clear that checklists vary considerably on this dimension. For instance, the Inventory of Small Events (Zautra, Guarnaccia, & B. P. Dohrenwend, 1986) was developed to be relatively objective, i.e., not reflective of internal states, whereas the Hassles Scale (Kanner et al., 1981) is more subjective in that feelings and other private states are tapped. One aspect of the discussion is the possibility that internal states (mood, depression, anxiety, etc.) influence the report of events that are subjective in nature (e.g., a depressed participant is more likely to report "problems"). A second aspect is that many minor events can be viewed as symptoms or as part of a clinical syndrome (e.g., irritability in depression [a symptom] expressed as arguments with family members [a minor event]). Because strong, yet meaningless, associations emerge when symptom-like events are used to predict symptoms, some researchers have recommended against the use of subjective events (B. S. Dohrenwend & B. P. Dohrenwend, 1974).

Many daily event studies of the type found in this special issue are prospective designs in which events are measured prior to the onset of the outcome, eliminating the second form of bias. However, if the event recording period is long enough to include the onset of the outcome, then bias is possible. For instance, if shifts in mood are the outcome variable and the event recording period is 24 hours, then biasing of event report is possible when same-day associations between events and mood are examined. Concurrent relationships of this sort are commonly found in the daily events and ESM literature.

Event reports should probably be as objective as possible to maintain a relatively pure, uncontaminated environmental measure. Appraisals of events (see below) can be used to address subjective qualities of the events. While it may appear that open-ended formats encourage reporting of more subjective events and close-ended formats favor more objective reporting, this may be illusory. In both modes of response there is ample opportunity for subjective events to be reported, because it is difficult to specify the boundaries of an event. It is not clear when a heated discussion (not included as part of a checklist) becomes the minor event "argument with spouse"; a variety of subjective con-

siderations (a person's past history with the event, current mood, etc.) probably influence whether the occurrence is reported as an event or not. The point is that close- and open-ended response formats are both open to this problem.

Summarizing event content. If extensive event sampling is being used to generate a checklist, the next step is to summarize the content. How this step is completed depends on the type of checklist desired. A theoretical approach to checklist development is to examine events that fall into the themes being explored, for instance, loss or overload. Events with these qualities might be identified through judges' ratings of events. Alternatively, if the object of the development is to characterize daily occurrences more generally, a content analysis approach, where content themes are allowed to emerge, may be employed (for an example of the latter approach, see Stone & Neale, 1982). This is an important step since it determines which events will be distinguishable from one another and which will be combined. It is also a very difficult step in the process because there is no clear theory to guide the effort. Furthermore, there has been relatively little discussion of this task in articles presenting new checklists.

As part of the development procedure, new events should be classified with a newly developed list to determine if there are ambiguities or overlapping events in the checklist. If a "test" event can be coded in more than one event category, then bias will be introduced into the measure either as an inflation in the number of events or as a misrepresentation of the event content. In this case, clarification of checklist items is required.

ESM studies have employed open-ended questions that are coded by the participants themselves (e.g., Brandstatter, 1983), by other participants (e.g., Campbell et al., 1991), or by trained coders (e.g., Wong & Csikszentmihalyi, 1991). Csikszentmihalyi and his colleagues have reliably coded responses to the question "What were you doing?" into 154 specific activity categories and 16 broader categories (e.g., work, leisure). Often, however, it is not clear how the categories of classification were derived, and this may have implications for the magnitude of the associations between event classes and outcomes.

It is clear that the development procedures, event sampling and summarizing of content, are critical to the nature of the resulting checklist. Unfortunately, very little is known about the best ways to proceed with

these steps, and future research should explore the effects of different developmental procedures on checklist content and the associations with outcome measures yielded by checklists developed with different methods.

Appraisal dimensions. Following the SRE strategy of assigning weights to experienced events, daily event inventories often gather information about the quality of experienced events. Event appraisals figure importantly into several stress and coping theories, perhaps most notably the transactional theory of Lazarus and his colleagues (Lazarus & Folkman, 1984), yet appraisal information complicates the interpretation of checklist data when event occurrence is weighted by appraisals. Appraisal questions are intended to enhance the predictive ability of the event information by focusing specifically on the psychological qualities of the event that are hypothesized to relate to outcomes and by allowing individual differences to emerge. In the ESM, most investigators have included subjective ratings of the situation at the time of responding. These appraisal dimensions have included challenge, skill, motivation, concentration, creativity, satisfaction, and relaxation (Csikszentmihalyi & LeFevre, 1989) and physical activity, arousal, tiredness, self-esteem, and sociability (Diener & Larsen, 1984; Diener et al., 1984).

Although the major life event literature suggests that weighting of events with group weighting coefficients does not improve prediction (Shrout, 1981), there is evidence that participants' personal ratings of events, made while completing the checklist, do improve prediction of outcomes (Sarason, Johnson, & Siegel, 1978).

The problem with using appraisals is the same as the one mentioned in the discussion of event objectivity. Subjective appraisals are prone to the influence of the outcomes studied. For example, being depressed may influence the appraisal of an event (increasing the degree of negativeness), resulting in a confounded same-day association. There are several ways to avoid or reduce this confounding in certain situations. For example, if event appraisal can be measured prior to the onset of the outcome, then the predictive associations will not be confounded since the outcome could not influence appraisals. Second, on logical grounds there are some appraisals that appear less prone to confounding by outcomes. For instance, it is difficult to imagine that an appraisal of whether the event was expected or not would be systematically biased by outcome measures such as mood or health. Another strategy for unconfounding appraisals and outcomes is to evaluate the meaning of

hypothetical events likely to be experienced during the study at the outset, before the outcome is present. Appraisals could then be used to weight subsequently collected daily event data. Since these appraisals are measured prior to outcome, they cannot be biased. On the other hand, these imagined appraisals are limited by the lack of information about the immediate context of event occurrence (e.g., other undesirable events occurring on the same day that could potentiate the impact of the event).

While strategies of this sort can often be effective in avoiding or limiting the confounding of appraisals with the outcome measure of interest, there are many other cases where this is impossible. Daily event researchers have not given much attention to this problem and it remains an important issue for future methodological investigation.

Although there has been no comprehensive study of event appraisals in daily checklists, there are unpublished data which address the predictive ability of several appraisals. The question of what is to be used as the outcome in such predictions is, of course, important: It is plausible that there are different predictors of mood, psychiatric symptoms, physical symptoms, and any of many physiological indices. In the analyses described above, prediction of same-day mood from appraisal dimensions was used for two reasons. First, it is difficult to imagine that health or even physiological processes would be affected without mood also being affected. Second, from a practical point of view, mood is measured every day on a continuous scale (unlike health measures which have relatively little variability in healthy populations) and, thus, is a reasonable outcome variable. Concurrent moods were predicted (negative and positive) separately from the following four event appraisals: desirability/undesirability, changing/stabilizing (to one's life-style), meaningfulness, and control (over the event's occurrence). Cross-products of various appraisals were also computed to determine if interactions of event appraisals predicted mood. In brief, we found that the desirability appraisal accounted for most of the variation in mood scores, and that the other dimensions added very little (Stone & Lennox, 1984). However, this finding may be an instance where event appraisals are confounded by same-day mood.

While it is clear that appraisals of daily events carry potentially important information about events' impact, it is notable that very little research has systematically explored what the relevant appraisal dimensions are for various outcomes. Unlike the work described above exploring the domain of events, comparable studies exploring the domain

of appraisals have not been conducted. It is conceivable that such analyses would discover that there are just a few dimensions that capture the essential psychological characteristics of daily events and these dimensions predict diverse outcomes. In other words, while there may be hundreds of daily events, it may be that they are important (for instance, stressful) for a limited number of reasons (for example, different kinds of threat). This intriguing and possibly very important issue deserves considerable future research.

Event validation/reliability. Validation of daily checklist data is extremely problematic, although it has been of surprisingly little concern to investigators. It is not clear that the usual kinds of validity, namely, discriminant, convergent, construct, etc., apply to event recordings. Events are not tapping underlying constructs such as anxiety, depression, or sense of control, where multiple indicators are used to hone in on the construct. Events are instead more comparable to behavioral recordings, where the event in question either did or did not occur: For example, the child either did or did not act out in class. A large set of items asking about a particular instance of acting out is not necessary here—there is no underlying construct to validate. Event occurrence is not supposed to represent or be a proxy for something else (e.g., anxiety).

On the other hand, reliability of event reports is extraordinarily important. A study's internal validity will be compromised if events are not reliably reported. As with other forms of behavioral observation, the meaning of event reliability is somewhat different than in the assessment of psychological constructs. It makes little sense to compute test-retest reliabilities since the critical question is whether a participant is reporting an event that actually occurred or not (a stable misreporting of events has little meaning).

As with other forms of behavioral observation, the idea that checklist items (events) should be internally consistent does not seem reasonable. This would imply that the experience of one event is related to other events (for there to be reasonable internal consistency), yet there is no reason to assume this. Again, assessments of constructs with multiple indicators do assume associations among items, but this isn't the case with event checklists.

But how is the researcher to obtain the "gold standard" occurrence information for comparison with participant reports of event occurrence

to evaluate validity? Securing consensus among various raters that the event actually occurred is one method to achieve this goal. The analogy in major life event research is some objective "proof" that a reported event (say the death of a family member or the loss of a job) actually happened. This is accomplished by viewing archival records or by interviewing significant others about the event. A few investigators in the major life events field who have used interview methodologies have been successful in verifying events (e.g., Brown & Harris, 1978), but for most checklist assessments this consensual validation is impractical.

Perhaps daily event researchers are less sensitive to the reliability issue than their major event counterparts because the former group feels that participants are eminently capable of performing their tasks. That is, they assume that the reporting of events for the last 24 hours is considerably less prone to retrospective recall biases and distortions than reporting major events for the past 12 months. This assumption should not be taken lightly. After all, although the recording period for events is much shorter in the daily model, the duration of the events being recorded is also much shorter and daily events are likely to be as prone to distortion and forgetting as are major events in the context they are measured. Since ESM researchers usually request that participants report their experiences immediately when signaled, recall biases are assumed to be minimal when participants respond in a short time. However, the issues of distortion and the accuracy of self-report have not been addressed in the ESM literature through the collection of observational data, although this strategy has been recommended (Hormuth, 1986).

Some data on this issue were provided from a study of wives' reports of their husbands' events. Husbands were the targeted participants in the study and wives served as observers of the husbands in an effort to obtain consensual validation of event occurrence (Stone & Neale, 1982). The procedure is admittedly imperfect because wives do not have access to all of their husbands' activities. Nonetheless, using a telephone interview procedure in addition to daily event recording, it was found that husbands did misreport many daily events: There were errors of omission (forgetting some event of the day), of commission (reporting yesterday's event today), and of simple disagreement about whether an occurrence should be recorded as an event or not. Thus, error-free recording of daily events certainly should not be assumed. One methodology that can be used to improve event assessment is to have both spouses discuss and agree upon what actually happened to

the target participant during a day in an attempt to achieve a consensual validation of occurrence. This approach is only applicable to certain situations where significant others are available and willing to participate in the research. It also adds an additional level of effort required by participants.

The reliability and validation of information obtained through the ESM have been difficult to establish. A number of factors related to the methodology are thought to affect the quality of the data. These factors include delayed, or missed, responses to signals; disruption caused by the monitoring procedure and the difficulty of responding reported by participants when they are debriefed; the possibility of inflated stability estimates due to adaptation to the signaling procedure and adaptation to repeated administration of the same questionnaire; and selective responding within certain situations and not others (Hormuth, 1986).

In summary, relatively little is known about event validity and reliability. As mentioned in Wheeler and Reis (1991), event researchers should not simply assume that their measures are valid and reliable. On the contrary, we have presented several reasons for exercising caution regarding these psychometric concepts.

Summary stress measures. Because most respondents report a large number of different events over a diary period, it is possible to develop much more complex multidimensional event measures in daily surveys than in surveys of major life events. Summary measures have been created in a number of different ways. One approach has been to code events within life domains (e.g., finances, work, family, health, etc.). This can be useful in research on stress spillover, for example, where the main research interest is in determining whether difficulties in one life domain create or exacerbate difficulties in other domains. It does not seem to be a good strategy, though, for studying the effects of daily events on health because event effects vary substantially within life domains as a function of more theoretically important dimensions. The latter can be captured more adequately either by creating summary measures for particular classes of events (e.g., arguments, overloads, role conflicts) or by creating summary measures from dimensional ratings (e.g., loss, threat, challenge). It is also possible to combine these latter two strategies, by beginning with summary measures of particular event classes (e.g., arguments) and introducing dimensional measures of the same events in an effort to determine whether the effects of the

classes on some outcomes can be explained as mediated through the dimensions (e.g., threat).

Designs and Considerations for Recording Events

In the following sections we discuss a number of considerations for designing and implementing daily event and experience studies. Our comments should be taken as recommendations, based both on statistical and methodologic considerations and on our experience in running daily event studies.

Determining sample size and number of observations

Statistical power considerations are as important in diary studies as in any other kind of hypothesis-testing research. Power considerations differ depending on the purpose of the diary analysis. In some cases, researchers use diaries as a way to obtain aggregate measures about some characteristic of participants that can be used as part of a cross-sectional analysis (Campbell et al., 1991) or as a baseline assessment in a two-wave panel analysis (Wong & Csikszentmihalyi, 1991). The diary reports, in this approach, are indicators of some underlying construct. The correct number of diary collections depends on the desired reliability of the construct. Wong and Csikszentmihalyi (1991), for example, used the ESM to measure how well students concentrate while studying, and the average respondent completed 35 Experience Sampling Forms. The estimate that studying occupies somewhat more than 10% of the student's day suggests an average of between three and four study episodes per student. Such a small sampling may not yield a reliable measure of typical concentration levels, assuming that these levels vary somewhat across study episodes. If this is true, then concentration while studying may actually be a more important determinant of academic achievement than suggested by Wong and Csikszentmihalyi's analyses, which assumed that the construct was measured without error.

This example illustrates a general point: that researchers who use the diary method to construct aggregate measures should apply the same standards for reliable measures as they would in constructing any other multi-item scale. The number of indicators required to achieve reliable measurement should be a central consideration.

Use of the diary design to generate information about desegregated person-time observations that are treated as the unit of analysis has been used by other researchers. Research by Larsen and Kasimatis (1991) and Bolger and Schilling (1991) are examples. The issue of appropriate sample size and the allocation of sample between participants and number of days is more complex in studies of this sort, because it requires a consideration of the joint sample of persons and times.

This issue can be approached on several levels. The simplest approach is based on the theory of statistical power for clustered probability samples (Cochran, 1963). The diary design is conceptualized as a single-stage cluster design of p clusters of Size d, where $p \times d = n$; for example, 100 persons $\times$ 30 diary days = 3,000 person-day observations. The problem of estimating the required sample size can be reduced to a calculation of the number of clusters (persons) over a fixed cluster size (number of diary days) needed to estimate parameters of interest with a prespecified level of power. Iterative estimation can be used to evaluate the trade-off between persons and days by calculating the minimum number of clusters for different values of fixed cluster size.

The subtlety in this conceptualization involves the fact that statistical power is not affected as much by changes in the size of d as the size of p. It is usually more expensive to add more people to the sample (increasing p) than to increase the length of the diary period (increasing d). Yet the more appropriate consideration is the improvement in power achieved by increasing either p or d by amounts of equivalent cost. Often, depending on the covariance structure of a particular data array, an increase of p by 1% will improve power more than increasing d by 5%, because day-to-day variation within persons is much less than between-person variation within a day. In evaluating the trade-off between p and d, it is also important to bear in mind that the quality of diary data is known to degrade as the diary period increases. An upper bound on d should be set in such a way as to avoid the more severe problems of declining data quality, which have been documented to occur between 2 and 4 weeks in some diary investigations.

Another consideration in determining the daily period (d) is the type of analyses that are required for testing a study's hypotheses. Shorter periods of recording are probably fine for creating reliable aggregated measures. However, some hypotheses demand much longer recording periods. Examples of such studies include those that are interested in

the effects of relatively infrequent daily events, such as a major argument between spouses, or those that require the occurrence of a particular outcome, such as an episode of respiratory symptoms. In these cases, the investigators may select participants in ways that increase the probability of the infrequent events (e.g., choosing marriages in distress or participants with histories of respiratory illness); however, some hypotheses may demand long study periods to achieve particular conditions. One additional consideration is the statistical analysis planned for hypothesis testing. Time series analyses usually demand many observations, especially if the lag periods being explored are long. This consideration is discussed elsewhere in this special issue (see West and Hepworth, 1991).

Event recording period

A second design issue involves the interval between waves of diary data collection, a concept related to the recording period interval mentioned earlier. While the interval between waves of collection is often the same as the recording period (as in the case of daily waves and a recording period of 24 hours), they need not be equal (waves separated by 2 months and a recording period of 1 week). The articles in this special issue illustrate the great variety found in the diary literature. Although the reader can see some rationale in the decisions regarding the recording period made by the different investigators, none of the articles included in this issue discusses the basis for the decision to collect data once a day, twice a day, or more often.

Two considerations in determining the spacing between waves are the length of time one considers theoretically important and the recording period over which the respondent is assumed to be able to report accurately. The first of these two considerations seems to be the main determinant in practice, with most diary researchers studying daily diaries (rather than twice-daily or thrice-daily, for example) because of a vague sense that there is something interesting or important about the day as the unit of analysis. This is a legitimate basis for determining the time interval if the purpose of analysis is to study repeated cross-sections rather than longitudinal associations. On the other hand, when dynamic analysis is the purpose for collecting time series diary data, the correct time interval hinges on some understanding of the true time lag between the variables that are to be included in the dynamic analysis.

In the absence of prior information to help document the correct time lag, the important dynamic effects can occur at a different time interval than the diary data collection.

The literature on macro time series analysis provides a number of procedures to help determine the correct time interval to capture dynamic processes. The most basic strategy is to plot cross-correlograms, which are graphs that describe the distribution of correlations between two variables (X, Y) over a wide range of leads $(Xt - Yt + t')$ and lags $(Xt + t' - Yt)$. Theoretical analyses of cross-correlograms have shown that diagnoses about lag structures can be made by studying the shape of these graphs (Kessler & Greenberg, 1981). In cases where we can rule out the possibility of Y causing X, for example, the peak value in the correlogram of lagged X is an upper bound on the true causal lag of that variable's effect on Y. This means that a diary analysis that seeks to capture the dynamics of this association must collect data at time intervals no longer than the interval of the peak value of the correlogram. More complex diagnostic procedures make use of cross-spectral analysis, which is mentioned by Larsen and Kasimatis (1991), to investigate a wide range of causal structures (e.g., distributed lags) and to determine the time intervals of the causal associations.

It is important to note that many daily investigations have not found significant lagged associations, but only concurrent associations. In the literature investigating daily events and mood, for example, most studies have observed and thoroughly characterized same-day associations; lagged associations are generally nonsignificant (Bolger, DeLongis, Kessler, & Schilling, 1989; Eckenrode, 1984). Although of interest, these concurrent associations are essentially very brief cross-sectional views of the processes under investigation. The advantages of the daily prospective design are lost in these analyses because the temporal precedence of the predictor variable is no longer assured.

An alternative approach to multiple daily assessments for the purpose of exploring within-day associations is to request that participants mark the time of event occurrence. If done accurately, this allows the researcher to use events that occurred in the morning to predict an outcome later in the day. This approach has been utilized by Haythornthwaite (1986) in the prediction of angina episodes from daily events. On the negative side, it remains to be demonstrated that participants can accurately note the timing of events. Observational studies of participants' behavior and responses to the timing questions are needed

before this approach can be used with total confidence. If it is shown that this approach yields data with good accuracy, it would be much less burdensome than completing two or three questionnaires for the same purpose.

Participant cooperation

Participant cooperation is a much more serious problem in diary studies than in most other types of research, given the considerable burden of completing diaries each day for an extended period of time. Special procedures have been developed to improve cooperation. We review the most promising of these strategies in this section.

Recruitment. Most diary research has relied on volunteers, in which case recruitment is not considered to be part of the cooperation problem. Volunteers, however, cannot provide information about the distribution of events or their typical effects in the general population and, for this reason, it is important that diary researchers develop methods to recruit representative samples of the general population. The literature on survey research (Groves, 1990; Traugott, Groves, & Lepkowski, 1987) provides strategies which can be used to recruit representative samples. These strategies need to be investigated and perhaps even extended to deal with the particularly severe nonresponse problems in diary studies.

Retention. Participation in a diary study requires the respondent to maintain interest in the study over a considerable period of time. This is more easily achieved when the diary is short, easy to complete, and enjoyable. A number of strategies have been used to help improve retention among diary respondents. One is to provide feedback on progress, either by periodic mailings with personal notes of thanks or by phone contacts. As noted earlier, reminders can also reduce nonresponse when respondents have failed to complete a diary at the scheduled time. Another strategy for use with daily checklists is to have the checklists mailed back at frequent intervals. For instance, daily checklists could be mailed back the day after they have been completed, which may motivate participants to complete them on the correct day. With weekly or longer collections of daily questionnaires, it is possible that participants may skip reporting for several days and then fill out several questionnaires at one sitting.

In a diary investigation recently carried out by Kessler and colleagues (Bolger et al., 1989), phone contacts with respondents who failed to return diary booklets indicated that they often decided to quit the study after inadvertently forgetting to complete their diaries for a day or two. They thought they had "ruined" their diaries and that it was no longer useful to the researchers to continue filling them out. A phone discussion with a member of the research team, informing the respondent that the researchers were very interested in obtaining diary responses for the full diary period even though there might be a break in the series, was able to convert most of these dropouts to resume participation in the study. This experience suggests that initial instructions to respondents should provide information that helps inoculate them from the inevitable setbacks that will occur in a large percentage of cases. Explicit instructions saying that completion of diaries over a specified number of days or over a specified calendar period is important whether or not it is possible to fill out the diary each and every day could be an important aid in capturing partial data from respondents who might otherwise become discouraged and drop out.

The experience of most diary researchers is that respondents who drop out of a diary study often do so during the first week. For this reason, as well as for reasons of providing rapid feedback on incomplete or inadequately detailed responses, researchers should build in special procedures for providing feedback and encouragement over the first week of the diary period.

The issue of material incentives is a complex one. Market research organizations that administer large consumer panels have all developed material incentive systems that allow respondents to choose among cash, gifts, points that can accumulate to earn large prizes, and sometimes even lottery tickets that provide some chance of receiving a very large prize. This multi-option system has evolved because marketers discovered that each incentive motivates some respondents, while none of the incentives motivates all respondents. Marketers provide comparatively more desirable prizes to those respondents who accumulate points, thus motivating respondents to remain in the panel for an extended period of time.

Although this experience in market research panels can provide some guidance to diary researchers, there are also ways in which the comparison breaks down. Most market research consumer panels have extremely low response rates, indicating that exclusively financial in-

centives in the range that is feasible for research of this sort are inadequate to motivate the vast majority of people in the general population. Furthermore, an exclusively financial incentive system runs the risk of attracting respondents who lack a commitment to honesty and accuracy in data reporting, further compromising research quality. Our own experience is consistent with this concern. One of our own studies involved a diary investigation of daily cold symptoms in a sample of college students who had been randomly assigned to one of several cold medications. Students who participated in the study were paid a substantial amount of money ($250) for two biomedical examinations (pre- and post-tests associated with the medication), regular use of the medication, and completion of daily diaries for a period of several weeks. There was great interest in this study, with many more students volunteering than were needed. Yet the quality of diary data was extremely low (e.g., high rates of missing data, evidence of all seven daily diary booklets being filled out on the same day, etc.). The substantial amount of money apparently attracted participants with no real interest in the diary completion task.

The challenge for diary researchers is to create a motivational system that encourages conscientious diary completion, not merely superficial participation for the purposes of material incentives. In an approach to the incentive problem consistent with the above points, Stone and colleagues have used monetary incentives for participation in diary studies, yet have made entry into the study relatively onerous by emphasizing the difficult nature of the collection task (e.g., frequent blood draws). The rationale for this approach is that participants attracted by large incentives or who do not realize the actual demands of the protocol may drop out quickly. Since entry into the study demanded rigorous at-home training by project personnel and weekly nurse visits for collection of biological samples, dropouts were especially costly. Thus, it was deemed better to emphasize the rigors of the protocol and not "overly" entice prospective participants with large incentives. Clearly, this strategy is not appropriate for epidemiological studies where a high response rate is important. Demographic and personality characteristics of those willing to participate in demanding protocols may be different from those not interested, and the findings from such participants may not generalize to nonrespondents.

Problems in Conducting Daily Event Studies

The problem of data quality has been addressed in terms of methods that can be used during the course of data collection. It is also possible to address this problem during the course of data processing.

Nonresponse bias

Research aimed at studying basic psychological processes may not need to be concerned about nonresponse bias, so long as the processes under investigation apply to the population in a way that is unrelated to the determinants of participation in the study. However, it often occurs that the research questions addressed in diary studies do not deal with fundamental processes of this sort, but rather with more basic descriptive questions that could be importantly affected by bias in participation.

There is little that can be done about nonresponse bias after data are collected unless the researcher has some information that can be used to compare participants from nonparticipants. This sort of information is often available in diary investigations that are based on general population samples, particularly if the diary sample was generated from a larger sample of respondents who participated in some sort of baseline survey. In cases of this sort, at least two broad strategies are available to the data analyst. These involve weighting and model-based correction for misspecification.

Weighting. Weighting for nonresponse can improve estimates if respondents differ from nonrespondents primarily in the distribution of characteristics that have been measured in the baseline survey. Although a number of different weighting schemes are possible and legitimate, one of the most appealing begins by estimating a probit equation, using data available from both respondents and nonrespondents to predict who participates in the study (Oh & Scheuren, 1983). Each respondent's predicted probability of participation based on this equation can be thought of as a summary representation of nonresponse bias for persons similar to him or her in the predesignated sample. For example, a respondent with a predicted probability of .25 can be thought of as the one person out of four with the same characteristics who participated in the study. A weighting scheme that adjusts for this nonresponse and, more important, for the differential probability of response of all participants, is one that weights this respondent and all others in the

sample by the inverse of their predicted probabilities of participation (e.g., $1/.25 = 4$).

Model-based correction for misspecification. Another approach is to work with unweighted data and introduce a control variable to adjust for nonresponse bias. Heckman (1979) has shown that the correct control variable for a linear regression equation or analysis of covariance model is the natural logarithm of the cumulative probability distribution of the predicted probability of participation, based on the same probit equation discussed in the previous subsection. An advantage of using the control variable approach rather than weighting is that a broader array of statistical analysis options are available in the analysis of an unweighted data array, including the ability to investigate whether critical predictor variables have effects on the outcome that vary with probability of participation in the study.

Attrition bias

Closely related to the problem of response bias is the problem of respondents who drop out partway through the study. This problem is somewhat easier to address than nonresponse bias because the researcher has some data on subjects lost through attrition. The situation is particularly favorable in diary studies, where the partial data usually consist of complete data records for an incomplete set of days.

The procedures available for addressing attrition are the same as those for nonresponse. Weighting, however, is a much more attractive option than model-based correction because complete data records for some diary days are usually available for dropouts. These data can be reweighted to reflect individual differences in days of participation in the study.

A critical assumption in this method is that attrition is not related to any variable of central importance to the investigation. If this assumption is incorrect, weighting will not have the desired effect. One could imagine the situation, for example, where the probability of attrition is related to the respondent's level of psychological functioning in relation to current event exposure, with a higher probability of rapid dropout for those respondents who are more distressed than one would expect from their event exposure. Regression coefficients predicting the selection variable (in this case, a scale of psychological distress) will

be biased toward zero by this type of attrition. Model-based correction procedures can remove this bias, but weighting cannot. Unfortunately, there is no accurate way to diagnose whether the functional form of the selection bias is like that specified by any particular model. Different forms of bias will lead to different types of error in data analysis and there is no guarantee that "correction" procedures will actually lead to improved estimates (Stolzenberg & Relles, 1990). The only safe option in a situation of this sort is to use several different estimation procedures and hope for convergence of results or bounding of critical parameters within a range of values that can be usefully interpreted.

Response decay

One important variant on attrition bias is the commonly observed pattern of response decay that is often found over the course of a diary study. This occurs when rates of reporting events go down over time for reasons that seem to be related to nothing other than the length of time the respondent has participated in the study. A pattern of this sort can reasonably be interpreted as reflecting the effects of fatigue, boredom, or measurement reactivity.

This problem can be dealt with in the same way as nonresponse and attrition—either by weighting or model-based correction. Another strategy is to exclude data from analysis once it becomes clear that decay of this sort is beginning to occur. Indeed, as noted in an earlier section, one of the most important determinants of the length of the study period should be evidence about this decay. The study should not go on for longer than a period of time for which it is known that decay does not occur unless there are theoretically important reasons that require a longer data collection period. In cases of the latter sort, length of time in the study can be used as a predictor variable in the weighting equation that is used to adjust for this special kind of attrition (Bolger et al., 1989).

Item-missing data bias

The types of missing data bias discussed up to now all involve case-missing data; that is, entire observations for one or more days that are missing from the data file. It is also common to find very high levels of item-missing data in self-report diaries. Item-missing data are missing

observations on individual questions on days when diaries are completed. In our experience, diaries are much more likely than one-shot questionnaires to have item-missing data, sometimes including entire pages that a respondent forgets to complete on a particular day.

The considerable body of research that has accumulated over the past few years on procedures for handling item-missing data will not be reviewed here. We will, however, make a few remarks about the central issues that have to be addressed by a researcher confronted with the high levels of item-missing data that appear in many diary studies.

As most readers are well aware, there are several standard procedures for dealing with item-missing data. These include limiting analysis to the subset of respondents with complete data (i.e., listwise deletion), using a pairwise missing covariance structure to fit aggregate models, using a data analysis procedure which allows for item-missing data, and developing item-level imputations prior to the beginning of data analysis which allow the full data array to be analyzed as if there were no item-missing data.

The first two of these options are the most commonly used in psychology. Neither is feasible in the analysis of daily diary data. Listwise deletion of missing values is infeasible because a very substantial proportion of respondents in most diary studies have some item-missing data on critical study variables. The subsample of respondents with complete data are likely to be unrepresentative of the total sample. They are also likely to constitute such a small number of cases that reliable estimation is jeopardized. Pairwise deletion is infeasible for a related reason; namely, that missing data are so common that pairwise associations can be based on substantially different respondents, running the risk of covariance matrices that are internally inconsistent.

These considerations argue for the use of thoughtful imputation and for reliance on data analysis techniques that allow for item-missing data. As noted above, there has been a great deal of recent statistical work on the imputation problem (Little & Rubin, 1987; Rubin, 1987). A wide array of strategies are now available for the researcher interested in using sophisticated methods for imputing item-missing data. Particularly attractive options exist in cases where multiple observations exist for a single respondent, as is the case in daily diary studies. Although not as yet introduced into the daily diary literature, we feel that these new methods for managing the serious problem of missing data should become a standard component in future diary investigations.

Linking event reports over time

Diary studies of daily events show that a substantial percentage of events that begin in a given day persist into the next day or even longer. For certain purposes, it can be useful to obtain information about linkages of this sort over time. Bolger et al. (1989), for example, showed that the impact of daily event stress on distressed mood varies depending on whether the event begins on the same day mood is being measured or has persisted over more than one day. This fact highlights the importance of viewing event information in a context and including strategies in the assessment to allow linking of information from one assessment to the next. The analytic considerations are especially salient since the meaning or appraisal of events may depend upon the context. A marital dispute that is reported on Day 25 may have different predictive ability depending upon whether it is the second or fifth day of a continuing argument or, alternatively, a 1-hour argument that occurred on that day only.

It is possible to obtain information about this persistence in telephone diary surveys, where the interviewer can review with the respondent the information he or she reported about stress on the previous day before asking about new events. It is more difficult to obtain information about linkages of daily events over time in self-report paper-and-pencil diaries, due to the logistic difficulties of asking respondents to review the previous day's events and record information about their termination or persistence. Nonetheless, checklists could include the question "Is this a continuation of the same event from the previous day?" which would be answered for all checked events.

The Future of Daily Event Studies

Several issues related to protocol development and study design deserve special attention in future diary studies. Regarding protocol development, researchers should pilot the use of a questionnaire, telephone interview, or the ESM. It is crucial that the researcher convince him/herself that the diary method adopted for the research is capable of assessing the constructs required for the research question. Furthermore, pretesting the method can provide information about the appropriateness of the event or experience content for the sample studied, about the task difficulty for participants, and about the attrition rate that can

be expected. These results may suggest that the researcher pare down or expand the instrument, rewrite certain items for the sample, change the recording period, and/or adjust the number of observations to be collected. Changes such as these can strongly influence the ultimate generalizability of the study's results.

Phone interview methods are generally superior to paper-and-pencil assessments, for the reasons mentioned above, and we strongly endorse the use of such techniques in future studies. Calls could be made once a day to achieve the same recording period as the daily questionnaires currently do, but calls made twice a day or more have the potential for producing an extremely rich set of data where within-day relationships and predictions could be examined. However, if telephone calling is not possible, we suggest that investigators design their questionnaires so that the respondents indicate the time of day that a checked event occurred. Information about event timing would allow within-day occurrences to be examined. Observed relationships could then be confirmed in the aggregate with more rigorous telephone or ESM interview methods.

Once the instrumentation has been chosen, the design of the study requires careful consideration. One starting point for this endeavor is to determine, either based on pilot results or on "best guesses," how often the event(s) of interest will occur. Not only the predictors (for example, events) but the outcome(s) must be considered. If event occurrence is infrequent, then relatively more recording days are required, whereas frequent event occurrences demand fewer recording days. The presumptive lag between predictor and outcome also must be considered. Clearly, the number of daily observations needs to be greater than the lag and, depending upon the frequency of predictor and outcome, perhaps many times longer if there is to be a reasonable probability of the predictor-outcome sequence to occur during the study. Finally, the researcher's opinion as to how long respondents will record accurately should be considered when determining how long the study will continue. As mentioned above, there is some evidence that event rates decline over a period of weeks of daily recording. It makes little sense to design a study that is much longer than participants' ability to provide valid data. Factors that can influence the quality of the recording were discussed above and should all be considered when deciding how many observations being recorded with what methods will best address a study's hypotheses.

Any form of the diary studies described above, ranging from the ESM to daily event recording, will entail a tremendous effort on the part of the research team. These studies involve considerable effort teaching participants the recording tasks, substantial effort in collecting the data and maintaining low attrition, many hours of arduous data entry and reduction, and complex statistical analyses to exploit the fullness of these data sets. These factors suggest that relatively few intensive diary studies will be run (at least compared to the number of cross-sectional and few panel studies), so researchers should be especially alert to the many potential opportunities that these studies may present. In other words, given the efforts that will be expended to complete the basic study, would a slightly increased effort yield considerably greater information by expanding the breadth of the design? Generally, the answer is yes. To explore individual difference in daily measures and associations among measures, researchers should carefully consider administering questionnaires and tasks at the outset of the study. Level of and variability of the daily measures is of much interest and may be predicted by stable personality and/or demographic factors. There are certainly other measures that could be completed on a daily basis without much additional burden to participants that could greatly enhance a study. For example, one might request that participants take their own oral temperatures by providing digital thermometers to confirm certain symptom reports (e.g., influenza). There are unlimited possibilities for additional data collection dictated, of course, by the goals of the research. Our point is that this issue is worthy of consideration.

CONCLUSION

A researcher is faced with a number of choices in the design of studies that assess daily events and experiences. There are important decisions to be made about the level of measurement of events required to allow appropriate testing of hypotheses, including the duration of the events to be measured, the period for which reports are made, the interval between measurements, the method used for recording events, and the domain of events covered in the assessment. Once these decisions have been made, the researcher must consider methods for sampling participants, for retaining participants once they have entered the study, for insuring high quality of daily data, and for keeping missing data to a minimum.

We have outlined our opinions about these and other issues in the daily events research area. As mentioned at the outset, there are no hard and fast rules about the conduct of event and experience research. We offer our comments as information that may be relevant to the researcher's decision-making process.

REFERENCES

Bolger, N., DeLongis, A., Kessler, R. C., & Schilling, E. A. (1989). Effects of daily stress on negative mood. *Journal of Personality and Social Psychology*, **57**, 808–818.

Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality*, **59**, 355–386.

Brandstatter, H. (1983). Emotional responses to other persons in everyday life situations. *Journal of Social and Personality Psychology*, **45**, 871–883.

Brown, G. W., & Harris, T. (1978). *Social origins of depression: A study of psychiatric disorder in women*. New York: Wiley.

Campbell, J. D., Chew, B., & Scratchley, L. S. (1991). Cognitive and emotional reactions to daily events: The effects of self-esteem and self-complexity. *Journal of Personality*, **59**, 473–505.

Cochran, W. G. (1963). *Sampling techniques*. New York: Wiley.

Csikszentmihalyi, M., & Larson, R. E. (1987). Validity and reliability of the experience sampling method. *Journal of Nervous and Mental Diseases*, **175**, 526–536.

Csikszentmihalyi, M., & LeFevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, **56**, 815–822.

Diener, E., & Larsen, R. (1984). Temporal stability and cross-sectional consistency of affective, behavioral, and cognitive response. *Journal of Personality and Social Psychology*, **47**, 871–883.

Diener, E., Larsen, R., & Emmons, R. (1984). Person × situation interactions: Choice of situations and congruence response models. *Journal of Personality and Social Psychology*, **47**, 580–592.

Dohrenwend, B. S., & Dohrenwend, B. P. (Eds.). (1974). *Stressful life events: Their nature and effects*. New York: Wiley.

Eckenrode, J. (1984). Impact of chronic and acute stressors on daily reports of mood. *Journal of Personality and Social Psychology*, **40**, 907–918.

Emmons, R. A. (1991). Personal strivings, daily life events, and psychological and physical well-being. *Journal of Personality*, **59**, 453–472.

Groves, R. M. (1990). *Survey costs and survey errors*. New York: Wiley.

Groves, R. M., & Magilavy, L. J. (1981). Increasing response rates to telephone surveys: A door in the face or foot in the door? *Public Opinion Quarterly*, **44**, 346–358.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research*, **14**, 121–132.

Hormuth, S. (1986). The sampling of experiences *in situ*. *Journal of Personality*, **54**, 262–293.

Haythornthwaite, J. (1986). *Daily life experiences and episodes of angina*. Unpublished doctoral dissertation, SUNY–Stony Brook, Stony Brook, NY.

Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. (1981). Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, **4**, 1–39.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of qualitative change*. New York: Academic Press.

Larsen, R. J., & Kasimatis, M. (1991). Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality*, **59**, 387–423.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer-Verlag.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

MacPhillamy, D., & Lewinsohn, P. M. (1975). *The Pleasant Events Schedule*. (Available from the Department of Psychology, University of Oregon, Eugene).

Oh, H. L., & Scheuren, F. E. (1983). Weighting adjustment for unit nonresponse. In W. Madow, I. Olkin, & D. Rubin (Eds.), *Incomplete data in sample surveys* (Vol. 2, pp. 143–184). New York: Academic Press.

Paty, J., Shiffman, S., & Kassel, J. (in press). Assessing stimulus control of smoking: The importance of base rates. In M. DeVries (Ed.), *The experience of psychopathology*. Cambridge: Cambridge University Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sarason, I. G., Johnson, J. H., & Siegel, J. M. (1978). Assessing the impact of life changes: Development of the life experiences survey. *Journal of Clinical and Consulting Psychology*, **46**, 932–946.

Shrout, P. E. (1981). Scaling of stressful life events. In B. S. Dohrenwend & B. P. Dohrenwend (Eds.), *Stressful life events: Their nature and effects* (pp. 29–47). New York: Wiley.

Stolzenberg, R. A., & Relles, D. A. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research*, **18**, 395–415.

Stone, A. A. & Lennox, S. (1984). *Prediction of daily mood from event appraisals*. Unpublished manuscript.

Stone, A. A., & Neale, J. M. (1982). Development of a methodology for assessing daily experiences. In A. Baum & J. E. Singer (Eds.), *Advances in environmental psychology: Environment and health* (Vol. 4, pp. 49–83). Hillsdale, NJ: Lawrence Erlbaum.

Traugott, M. W., Groves, R. M., & Lepkowski, J. M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, **51**, 522–539.

Verbrugge, L. M. (1980). Health diaries. *Medical Care*, **18**, 73–95.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, **59**, 609–622.

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, **59**, 339–354.

Wong, M. M., & Csikszentmihalyi, M. (1991). Motivation and academic achievement: The effects of personality traits and the quality of experience. *Journal of Personality*, **59**, 539–574.

Zautra, A. J., Finch, J. F., Reich, J. W., & Guarnaccia, C. A. (1991). Predicting the everyday life events of older adults. *Journal of Personality*, **59**, 507–538.

Zautra, A. J., Guarnaccia, C. A., & Dohrenwend, B. P. (1986). Measuring small life events. *American Journal of Community Psychology*, **14**, 629–655.

Manuscript received July 9, 1990; revised November 29, 1990.