Chapter 8

Ecological Validity for Patient Reported Outcomes

Arthur A. Stone and Saul S. Shiffman

Asking people about their health, symptoms, attitudes, opinions, and behaviors is ubiquitous in the behavioral, social, and medical sciences (Stone et al, 2000). For many areas of inquiry in these fields, it is impossible to contemplate research programs without self-reports. Self-reports often serve as primary outcome measures, for instance, in assessing pain, fatigue, opinions, or attitudes; self-reports are the accepted standard for these constructs and "objective" alternative measures usually are not available. Even when objective measures are possible in principle, we often rely on self-report data (e.g., smoking behavior, asthma attacks, social interactions), because the costs of objective data collection (via behavioral observations, for example) are prohibitive.

Patient Reported Outcome (PRO) is a new term used to describe self-reports when they are used as outcome measures in trials (FDA Docket No. 2006D-0044; Rock et al, 2007). The importance of PROs to the behavioral and medical research enterprise has been highlighted recently. The US Food and Drug Administration is in the process of setting standards for PROs used in clinical trials submitted in support of drug or device approvals and claims (FDA Docket No. 2006D-0044). The National Institutes of Health (NIH) has also devoted one of its Roadmap Projects, which are large-scale, high priority initiatives intended to advance health research, to the development of psychometrically sophisticated PROs for use with chronically ill individuals participating in clinical trials (www.nihpromis.org). There is also no doubt that PROs are essential for the delivery of medical care, where they provide essential information about patient functioning and satisfaction with services.

An important feature of PRO assessments, as they have traditionally been implemented, is that they have generally been obtained in relatively artificial or unusual settings, such as clinics and research laboratories, and by having participants recollect and/or reflect on their past experiences. The purpose of this article is to discuss the potential value of collecting PRO data in participants' natural environments and with minimal recourse to recall by systematically and repeatedly sampling self-reports in peoples' daily environments, offering the possibility of truly representative sampling. In the first section of the article, we review the concept of sampling everyday life, its implications for ecological validity, and how it could affect self-report information and PROs. We discuss studies from cognitive science, autobiographic memory, and survey design inform this

A.A. Stone (✉)
Department of Psychiatry and Behavioral Science, Stony Brook University, Stony Brook, NY 11994-8790, USA
e-mail: arthur.stone@sunysb.edu

AAS is the associate chair of the Scientific Advisory Board of invivodata, inc., a company that supplies electronic data capture services for clinical research and is a senior scientist at the Gallup Organization. SS is a founder of invivodata, inc. and the chair of its Scientific Advisory Board.

A. Steptoe (ed.), *Handbook of Behavioral Medicine*, DOI 10.1007/978-0-387-09488-5_8,

discussion. In the second section, we review methodologies and technologies that enable collection of self-reports in peoples' typical environments, enhancing the representativeness of the resulting data.

1 Ecological Validity and Self-Reports

Today, when we think of the degree to which behavior observed in a research setting such as a research laboratory is generalizable to real-world behavior, we call this "ecological validity" (Hammond and Stewart, 2001). Over 70 years ago, ecological validity was first used in Brunswik's 1944 paper examining the perceptual phenomenon known as size constancy (Brunswik, 1944) – the ability of people to correctly judge the size of objects despite the fact that the projection of objects on the retina varies with viewing distance. Brunswik's interest was in how naturally occurring cues associated with objects, such as distance from object, were used by the individual to estimate size. In one study that presages the methods described later in this chapter, he recorded over several weeks randomly selected moments from a subject's daily routine and noted the retinal projection (via a photograph of the object), object size, and the subject's estimation of size.

The innovative feature of his design was the evaluation of the natural, ecological association of objects and their associated cues, in contrast to possibly artificial associations based on laboratory investigations, where the constellation of stimulus qualities bore little resemblance to those encountered by people in everyday life. "Representative design" was the term Brunswik coined to refer to the degree that a laboratory experiment corresponded with a particular set of environmental circumstances to which the results of the experiment were to be generalized – what we now call ecological validity. In keeping with contemporary parlance, we use the term ecological validity in its modern meaning, while acknowledging its historical evolution.

2 Momentary, Retrospective, and Global Self-Report

For this discussion, we describe three types of self-reports defined by the cognitive tasks inherent in making the reports. We shall refer to these as *momentary states*, *retrospective summaries*, and *global reports*. Momentary state questions ask people to describe some aspect of their immediate state, for example, their current mood, symptoms, and circumstances. A question about immediate pain intensity could use, for example, the following wording: "Please indicate your *current* pain intensity."

Most assessment in medicine and behavioral science, however, does not focus on momentary assessment, but for practical reasons typically asks for a summary of experience over a period of time or about a past experience at a particular time. These are called retrospective self-reports, and the time frame for these questions can range from the last day to one's entire life. The important idea is that the intention of the question is to capture information outside of immediate experience, which is presumed to be available in memory. Examples of typical recall questions include "Please indicate your average pain intensity over the past month," which asks the respondent not only to recall but then to summarize (average) the retrieved results, and "When was the last time you stayed overnight in a hospital?" which asks for a specific fact relating to a particular occasion.

The third type of self-report, global report, does not have any time frame at all, but rather asks the respondent to generalize globally or universally. "Generally speaking, how happy a person are you?" and "Are you prone to anxiety?" are examples of global questions. These questions seek information about a person in general. They might be equivalent to retrospective summaries over a lifetime, but that is not clear.

3 Does Ecological Validity Matter for Self-Report?

Our focus is on the relevance of ecological validity in the three kinds of self-reports, and we believe this depends on two things. The first is whether or not the phenomenon to be captured by self-report varies over time and situation, and the second is whether or not individuals can accurately recall and summarize it without distortion. In brief, we believe that special procedures are needed to assure ecological validity when a phenomenon varies over time *and* when respondents are not able to accurately recall and/or summarize it. Under these conditions, asking respondents for their impression of experience over some finite time period will yield results that may not accurately reflect real-world experience.

3.1 Variability over Time and Situation

When the variable under study does not vary with time and circumstances (e.g., the respondent's gender), any method of self-report (your current state, your state yesterday, your state in general) will yield the same answer, making issues of recall and ecological validity moot. However, most of the phenomena we study do vary over time for several reasons. They may vary due to the impact of the immediate context (physical setting [work/home, outside/inside, and other physical qualities] or social setting [whom with, type of activity, and other interpersonal qualities]); due to maturation of the individual and associated change; or due to temporal effects such as time-of-day, day-of-week, season, etc. These factors create true variation, not just variation due to measurement error (noise), and investigators have an interest in that true variation. When such variation exists, and individuals are not capable of producing an unbiased summary of the variable experience (discussed in next section), consideration of ecological validity is essential.

As an example of how environmental variability demands consideration in the design of studies to ensure ecological validity, consider an investigator who is trying to characterize participants' emotional state over a period of time. Now, affect is known to vary depending on the circumstances and setting. To achieve an accurate assessment of "average mood," which might serve as an investigator's outcome variable in trial, one would need to consider the full range of settings that the individual encountered – their mood may have been relaxed at home, but tense at work, or relaxed at work on Tuesday, but tense at work on Wednesday. In this case, it would be misleading to assess mood at work only or on Tuesdays only. The full range (or a representative range) of experiences and contexts would need to be taken into account, and properly weighted, to achieve an unbiased assessment of mood over the period. If one believes that individuals are capable of retrieving this information and weighting it appropriately, then recall summaries would be considered valid. If one concludes that we are not consistently capable of such cognitive feats, then one may need to actually sample and assess experiences across a range of time and settings (methods for doing so are discussed later).

3.2 Accuracy of Recall and Summary Processes

We have suggested that conclusions about respondents' ability to accurately recall and summarize the past are vital to determining how one collects data. Key to appreciating the limits of autobiographical memory is understanding the process of recording, retrieval, and summary of information about past experience. How, then, do we generate recall of and summarize our past states? Research indicates that the process of generating such "memories" is more accurately characterized as *reconstruction* (Menon

and Yorkston, 2000; Schwarz and Strack, 1991) than simple retrieval.[1] Memories can be reconstructed using a variety of heuristic strategies to build plausible responses that usually serve adequately for memory's everyday adaptive uses. The use of heuristics is a critical point, because heuristic strategies can introduce significant bias. Ironically, for environmentally sensitive variables, the subject's state *at the time of the recall*, which is itself subject to the effects of the recall setting, can influence recall and summary processes. For example, several studies have shown that the pain levels experienced at the time of assessment biases the recall of past pain, such that respondents in current high levels of pain recall more pain (Eich et al, 1985; Linton and Melin, 1982; Smith and Safer, 1993).[2] In another example, Schwarz has shown that very small pleasures (finding a dime) just prior to assessment can have large impacts on responses to global well-being questions (Schwarz, 1996). Or, that bringing to mind remembrances of events that pertain, at least in part, to the broad question have the effect biasing responses toward the recently recalled experiences (Schwarz, 1996). Current states skew both what information we retrieve about the past (e.g., mood congruent recall; Clark and Teasdale, 1982) and how we interpret that information. In other words, our summaries of past experience are not built from objective, statistical summaries of the past, but are influenced by our present condition.

In a similar way, participants' recall of experience is overly influenced by the most intense and the most recent experiences during the target reporting period; this has been called the "peak-end" effect (Fredrickson, 2000). Both the undue influence of our current state and that of recent and intense experiences are attributable to the influence of what is most "memorable" or salient, and the consequent under-weighting of routine experience – the fabric of everyday life – often resulting in systematic bias in recall (Kahneman et al, 1999). It should be noted that these heuristics operate rapidly and out of consciousness, as demonstrated by their impact in laboratory studies examining short-term recall (e.g., Redelmeier's colonscopy studies, Redelmeier and Kahneman, 1996). So, research participants, who are usually doing their best to provide accurate recall, are not aware that their recall reports are biased and have no ready way to avoid the bias. Not only do heuristics produce bias (that is, systematic errors) in contrast to merely injecting "noise" (random error) into recall, but the use of particular heuristics may vary between persons and across contexts, making it difficult to devise strategies that correct for heuristic bias and, more broadly, making the interpretation of recall reports exceedingly challenging.

Recall is also influenced by semantic memory, that is, generalized knowledge or belief (e.g., about myself, about work; Robinson and Clore, 2002; Ross, 1989). This may be especially prevalent when memories of an event, which may or may not be accurate, do not spring into mind. Memories constructed in this way are often "adjusted" to make them conform to logical scripts about events based on broader beliefs about behavior (in general or one's own) – they represent "what should have happened" or "what must have happened." Ross (1989) has shown that participants distort their recall to conform to their "personal theories" about behavior, for example, ideas about how stable or changeable their behavior is, beliefs about the influence of events on behavior, or their beliefs or ideals about themselves. These biases are particularly troubling, because they can generate "recall" that is psychologically coherent and consistent with theory (and thus easily accepted by scientists), but not based on fact. For example, participants who believe they have painful menstrual

[1] In fact, retrieval is not a simple process in that what is retrieved may be influenced by the individual's psychological state at the time of retrieval. For example, unpleasant memories are more accessible when an individual is in a negative affective state than when in positive affective state (Kihlstrom, et al 2000).

[2] A respondent's affective or pain state at the time of retrospection also influences the accessibility of certain memories and the heuristic processes used to summarize retrieved memories.

periods tend to "recall" such pain (and investigators may accept such reports), even when their own real-time reports showed they did not experience them (McFarland et al, 1989; see also Shiffman et al, 1997). Thus, cognitive science tells us that autobiographical memory is subject to substantial biases that can significantly distort self-reports. We next examine the implications of recall and summary processes for the different types of self-reports.

3.3 Implications for Global Reports

Evaluating the impact of accuracy and summary processes on global reports is difficult because it is not clear exactly how global assessment *should* line up with actual experience. If one assumes that global questions are meant to or are interpreted as reflecting experience – perhaps not an unreasonable assumption in many cases – then all of the troublesome processes associated with recall reports are applicable. Furthermore, there is evidence that ambiguity about what information is sought by a question and/or the inability to access that information from memory disposes respondents to answer on the basis of semantic memory (Robinson and Clore, 2002). Particularly when it is not clear what memories are relevant over what period, global questions will tend to pull for answers based on beliefs and attitudes. Although semantic memory has a connection to experience, that connection can be a loose one because other factors, such as beliefs, personality, and contextual cues. If it is actual experience that one seeks, then answers based on semantic memory are not ideal.

On the other hand, if one is interested in beliefs or opinions – and not actual experience – then global reports may be optimal. Beliefs and opinions can shape current and future behavior, so are of practical value and worthy of study in their own right, but care must be taken to distinguish between these beliefs and actual past behavior and experience, which may not be accurately reflected.

3.4 Implications for Retrospective Reports

The validity of retrospective self-reports depends on reporters' ability to accurately recall experience. As discussed above, cognitive research suggests that much of the information we seek about past behavior or experiences is not available in memory; we simply do not store such detailed and comprehensive information (Bradburn et al, 1987; Robinson and Clore, 2002; Schwarz and Oyserman, 2001; Schwarz et al, 1994; Thompson et al, 1996). Accuracy of recall and summary processes are, then, a major concern for interpreting recall reports. The extent of the concern, however, should be moderated by the nature of the recall content, as certain material (e.g., major, "unforgettable" events) may be less susceptible to memory failures, although still may be subject to the vagaries of summary processes.

It is also important to recognize that even "incorrect" or distorted recall can have substantial predictive validity. Some studies have shown that one's distorted memory or characterization of events can be a better predictor of future behavior than the actual experience. After all, it is this stored summary, however biased, that we later retrieve as a reference for future informing attitudes or directing our behavior (e.g., recalling how painful a previous colonoscopy was in order to decide whether to get another one; Redelmeier et al, 2003). Thus, there is value to the information held in retrospective reports, even when it does not faithfully reflect experience, but care must be taken not to interpret it as a true account of past events.

3.5 Implications for Momentary Reports

Assessments of current experience are not subject to recall bias, so the heuristics associated with memory processes are not much of a problem for these assessments. In contrast to the

difficulty recovering accurate information about the past, Robinson and Clore (2002) and others have argued that we have good access to our current or very recent states; that is, questions about immediate state are answered by retrieval of experience and not by reference to beliefs. However, to say that momentary reports are immune to recall biases is not to say that such reports are entirely accurate and reliable, because self-reports are susceptible to other distortions that can influence the assessment (for example, the desire to present oneself in a favorable light; Schwarz, 2007), but at least the biases introduced by memory processes are minimized.

In summary, recall and global questions are prone to bias due to the limitations of memory capacity and to the ways that people reconstruct and summarize experiences over time. These biases threaten the validity of the resulting reports when those reports are meant to represent the actual experiences the individual had over the specified period recalled. Immediate reports can escape biases due to recall processes, but raise new challenges for achieving ecological validity.

4 Rationale for Taking Self-Report into Everyday Life

Despite the potential problems identified for recall and global questions, these types of questions have dominated the field of self-report assessment. First, recall is subjectively compelling: We trust our own memories unquestioningly most of the time, so it seems natural to trust our participants' memory as well, particularly when they don't seem to have a motive for dissembling. Yet research has shown that confidence in a particular memory is often unrelated to its accuracy (Busey et al, 2000; Wells and Bradfield, 1998). Additionally, the nature of memory and its tendency to bias is a relatively recent discovery and has not yet penetrated deeply into thinking about research methods.

Recall methods are also used because they are enormously convenient and efficient: In a relatively brief period, the researcher or clinician is able to gather information on long periods of time, often up to years in duration, and on a wide variety of environments. If recall and global methods were capable of providing accurate information over such periods, there would be very little reason to consider alternatives. But, as the prior section of this chapter has shown, recall and recall self-reports may not be up to the task of providing truly accurate information about experience, at least some of the time.

If memory cannot be relied upon, then momentary assessments become essential. However, momentary assessments are limited by their very immediacy and narrow focus to what is happening *now*, at the moment of assessment, which is not often the investigator's focus. We earlier stated that many phenomena of interest vary across time and environmental context. It follows that momentary reports of those phenomena will vary by context. Thus, if momentary reports are to represent the person's overall experience, they would have to be collected in those contexts. No one momentary report could represent the subject's experience – there would have to be many. And, to achieve ecological validity, they would have to be collected in a wide range of real-world contexts, representatively sampling participants' momentary states across the range of settings they encounter.

These elements – real-time data collection about momentary states, repeated assessments, and sampling of real-world settings – form the core of the approach we have called Ecological Momentary Assessment (EMA; Stone et al, 1994, 2007; Shiffman et al, 2008).

Modern EMA methods have made use of innovations in data-collection technology, but EMA is not primarily a technological development. It more fundamentally addresses the design of data-collection protocols in relation to study objectives. Bolger and colleagues (2003) have enumerated three broad functions of EMA data collection: characterizing persons and individual differences (e.g., level of depressive symptoms); estimating within-person variability (e.g., standard deviation of pain intensity levels over a 1 month period); and estimating

within-person associations among two or more variables (e.g., association between changes in sleep and gastrointestinal symptoms the following day or between time-of-day and fatigue levels). The reader is referred to Stone et al (2006) and Shiffman et al (2009) for examples demonstrating these uses of EMA data.

Aside from addressing issues of recall bias, EMA methods conceptually address other issues discussed in the psychological assessment literature. The first concerns the arbitrary nature of measurement often associated with psychological assessments, a topic recently reviewed by Blanton and Jaccard (2006). In essence, it is difficult to understand the meaning of scale scores on many instruments, because they are not linked to other referents. So, when an individual moves from an affect score of 50–60 on a 100-point scale, it is impossible to know exactly how their affect has changed. Because EMA protocols can representatively sample over time, it is possible to express the observations by estimating the proportion of time an individual has experienced some state (e.g., is angry, by some definition) or is in a particular environment (e.g., at work). Such "prevalence" metrics offer the advantage of being easily interpretable and, further, they possess ratio level measurement qualities. The clear labeling afforded by such measures enhances the opportunity to develop strong theories and interventions, which is not the case when there is less certainty about a measure's meaning (Blanton and Jaccard, 2006). Also consistent with recommendations of Blanton and Jaccard is the emphasis on the assessment of real-world occurrences.

A second conceptual issue concerns the place of EMA data in an assessment model, which is pertinent to considerations of its usefulness in theory development. Here we refer to the framework developed by McFall (2005) in an article on theory and utility in evidence-based assessment. In our view, self-report EMA data can be considered an instance of a "sample" approach versus the alternative "sign" approach to assessment. This is because EMA measures often directly assess the target experience or behavior, rather than some other construct that is simply associated with the target. Signs, on the other hand, are indirect measures that simply have predictive utility (as in an actuarial prediction where any variable statistically associated with an outcome can be used to improve prediction, even if it has no conceptual overlap). Importantly, because there can be recall and summary problems with self-report data that can invalidate an assessment, EMA data may have unique value in providing proximal sample data for assessment. For example, one method for measuring coping with difficult events is based upon 1-month recall of the problem and the thoughts and behaviors used to cope with the problem. We compared real-time reports of these thoughts and behaviors with the recalled ones and found major discrepancies (Stone et al, 1998). Similarly, we compared global reports of smoking patterns to detailed real-time self-monitoring and found little correspondence (Shiffman, 1993); only the real-time data predicted subsequent relapse (Shiffman et al, 2007). In both cases, the real-time data might be considered a preferable sample.

The next issue concerns the isomorphism between recall measures of an outcome and EMA measures of an outcome. As mentioned above, using EMA to characterize overall levels of a self-report variable over a defined period of time is one of its primary uses. Little would be gained by using EMA methods if the resulting data were identical to those obtained by recall methods. Although there is a surfeit of information about *potential* reasons for achieving different results with the two methods, there is a paucity of empirical data documenting differences. In directly comparing data produced by the methods, two types of comparison emerge: (1) differences in levels (assuming the same measurement metric was used for both methods) and (2) differences in correspondence between rank-orderings of individuals by the methods (e.g., the correlation between the scores) (Stone et al, 2004; Shiffman et al, 2008).

Our own work on the assessment of pain intensity in patients with chronic pain disorders has partially addressed this question. Regarding differences in level of reporting, retrospective

assessments produce higher levels of pain when compared to the average of momentary reports for the same period of time, and the discrepancy between the methods increases as the reporting period increases (Broderick et al, 2008). One possible explanation for these results is that the peak heuristic, which posits a particular focus on high levels of past pain, leads to an overemphasis of bouts of pain in the recalled reports (Stone et al, 2004). Others have also observed the higher level of reporting with recall measures (Linton and Gotestam, 1983; vandenBrink et al, 2001). On correspondence between the two methods, the situation is less clear because although there is a substantial correlation between the pain reports from the two methods (about 50% of the variance is shared), there is also a substantial proportion of variance that is unique to each method. This general finding led earlier researchers to call it a "half-empty or half-full" situation, depending upon one's perspective (Salovey et al, 1993).

Whether or not the magnitude of the association seems acceptable, there is evidence that recalled reports can be distorted in undesirable ways. For example, we found that how much pain a respondent experienced at the time of reporting their retrospective weekly level influenced the magnitude of retrospective report (Broderick et al, 2006). We have also reported that the degree of variability in EMA pain reports over a week is associated with recall of pain over the same week (Stone et al, 2005). The degree and direction of differences between recalled and actual immediate experience, and how these are affected by study conditions, needs further empirical exploration.

5 Conducting EMA Studies

Our purpose in this section of the chapter is to provide the reader with an overview of the many issues that confront the researcher endeavoring to collect self-reports from everyday life. The presentation focuses on design considerations relating to ecological validity, but the reader is referred to many excellent comprehensive reviews (Affleck et al, 1999; Bolger et al, 2003; Delespaul, 1995; Shiffman et al, 2008; Stone et al, 2006). EMA is comprised of a variety of sampling designs that can be used singly or can be combined to meet the needs of investigators (Shiffman, 2006). A variety of schemes for scheduling assessments to ensure a representative sample of moments have been described (Delespaul, 1995; Shiffman, 2006).

The most commonly used schedules sample participants' experience through time-sampling; that is, they select a random sample of moments for assessment. The classic examples, from Experience Sampling Methodology (Csikszentmihalyi and Larsen, 1987; DeVries, 1987), are studies where participants are "beeped" at random times and prompted to complete an assessment of their momentary state. Random sampling of moments is seen as the key to representativeness, much as random sampling of individuals is seen as important for characterizing populations. As with sampling of individuals, any given sample of moments from a period of time will not yield a perfectly representative picture of a self-report variable; there will be an associated sampling error, just as there is when sampling people. Greater numbers of samples yield estimates with smaller sampling error.

Random time-sampling is not the only assessment schedule used in the EMA literature. An alternative is to schedule assessments at particular times of day, for example, every 2 h after 10 am, as a way to capture the day's experience. The limitations of this approach are discussed in Shiffman et al, (2008). Another alternative scheduling scheme is not based on time at all, but instead focuses on assessing particular events of interest. Thus, participants might be asked to complete an assessment every time they smoke a cigarette or engage in a social interaction. These event-based methods, which evolved from behavioral self-monitoring (Korotitsch and Nelson-Grey, 1999), are best suited to contexts

where the phenomenon of interest is a discrete event (e.g., an asthma attack) or can be construed into episodes (e.g., exacerbations of pain).

A few examples can help characterize EMA methodology: In one study, patients with rheumatologic disorders rated their pain and mood up to 12 times a day when prompted at random times by a computer to complete an assessment (Stone et al, 2004). In another study, problem drinkers tracked each episode of drinking, recorded their level of intoxication and how they felt about their drinking (Muraven et al, 2005). A third study assessed the symptoms of people complaining of multiple chemical sensitivity several times per day, while simultaneously sampling the surrounding air for analysis of chemical exposures (Saito et al, 2005). In a study illustrating a combination of time-based and event-based sampling, Shiffman and Waters (2004) used time-sampled data to examine trends in affect in the days and hours preceding a focal event (smoking relapse). While the subject populations, assessments, and content focus differed, these EMA studies and others (Stone et al, 2006) share an approach involving multiple momentary assessments, collected near the time of experience, across a broad range of real-world settings the participants inhabit, and with attention to sampling of experience (e.g., random time-sampling). These are the core elements of EMA.

In another parallel with sampling of research participants, EMA researchers have been concerned about the loss of observations from the planned sample and accordingly have emphasized the importance of compliance with scheduled assessments and inclusion of all relevant moments in the sampling frame as key to representativeness (Hufford and Shields, 2002; Stone et al, 2002; Shiffman et al, 2008). Just as attrition from a sample of participants threatens the representativeness of the sample, so non-compliance with assessment prompts threatens the representativeness of the sample of moments. A variety of EMA sampling schemes, paralleling the variety of sampling designs for individuals in populations, have been described and used (stratified sampling, over-sampling, etc.) (Shiffman, 2006).

5.1 Implementation of EMA and Application of Technology

Advances in technology have enabled the conduct of efficient and imaginative EMA studies. Early diary studies had no reliable way of scheduling assessments or prompting participants to complete them, so assessments were often linked to standard events in participants' lives, such as meals or bedtime. However, these are hardly random moments in a person's day. An innovation was introduced by the developers of the Experience Sampling Method, who provided participants with electronic pagers and arranged to "beep" them to prompt them to complete a diary card (Csikszentmihalyi and Larsen, 1987). By providing a means of signaling the subject, beepers gave the investigator control over the intended schedule of assessments, which were typically recorded on traditional paper diary cards.

The use of electronic data capture for EMA has become increasingly common. Besides scheduling and issuing prompts, a palmtop can also collect and store the assessment data, while recording the exact time the assessment was completed. This is regarded as an important advantage, because of concerns about back-filling of data – that is, the completion of assessments after-the-fact, with falsification of the completion date and time, which negates the advantages of real-time data collection. There has been controversy about how often back-filling occurs, how it might be minimized, and what effects back-filling has on the resulting data (Green et al, 2006).

Nevertheless, several studies, with diverse populations and methods have demonstrated that participants do back-fill paper diary entries, even when they are electronically prompted for completion, and sometimes even when they are

aware their entries are subject to verification (Hufford, 2007). This can be a serious concern, because participants who complete their diaries in retrospect reintroduce all of the problems of retrospective recall that the method was designed to avoid. Moreover, when participants choose when they complete the diaries, even if it's not long after the scheduled time, they can introduce additional bias because participants' choice of occasions can be biased (e.g., waiting until a symptom-free time to complete a diary or completing it when symptoms occur and serve as a reminder to do the diary). In essence, the sample of moments becomes like a convenience sample of volunteers, rather than like a random population sample. Accordingly, the ability of electronic data-collection methods to accurately record the time of diary completion is regarded by many investigators as an advantage over paper-and-pencil diaries.

Another advantage of many electronic data-collection systems is that they allow flexibility in the administration of questions, for example, item presentation can be contingent upon responses to prior items (e.g., skip patterns), greatly enhancing efficiency and reducing subject burden. Moreover, such electronic systems can also modify the sampling schedule based on algorithms applied to subject input, for instance, increasing the density of assessment when an event of interest has occurred or scheduling a series of assessments to follow up on a trigger event.

The most commonly used electronic devices for collecting self-report EMA data are palmtop computers and interactive voice response systems (IVRS). An advantage of palmtop computers is that they function independently and thus are not dependent on communication to a central center. They are also capable of presenting a variety of response options (Likert scales, Visual Analog Scale [VAS], Numeric Rating Scale, body diagrams) that are typically used in assessments. Since they present assessment content as text, the assessments resemble their paper ancestors, which probably accounts for the finding that such electronic assessments are psychometrically equivalent to parallel paper forms (Gwaltney et al, 2008).

In IVRS, assessment content is played to participants via recorded voice, and participants record their responses using the keypad ("press '1' if you are suicidal. . ."). While IVRS is most often used as a passive system requiring participants to call in, it can also be used to call participants on a schedule enabling time-sampling designs. With the advent of cell phones, the phone system can be used to reach participants in a wide variety of settings. An advantage is that IVRS uses the telephone – a technology familiar to participants. A disadvantage is that aural presentation of assessment and response options can limit the assessment (e.g., memory capacity limits the number of response options) and might change how participants respond.

As cell phones become more sophisticated, "smart phones" are increasingly able to function much like palmtop computers, displaying text-based assessments and sending assessment data to a central server. Desktop computer systems (web-based or otherwise), while not portable and thus not amenable to assessment in the full range of participants' settings, can be used to administer end-of-day or periodic assessments.

At the same time, these approaches are used to collect self-report data; a variety of specialized hardware can be used to assess participants' objective physiological states in a momentary way (e.g., ambulatory blood pressure, blood glucose, pulmonary function) (Kamarck et al, 1998). Other devices can objectively capture subject behavior (e.g., instrumented pill bottles, motion-detectors, audio or video recordings (Byerly et al, 2005)) or environmental conditions (e.g., noise, temperature, presence of chemical pollutants (e.g., Saito et al, 2005)). Collection of such objective data is often enriched by collecting concomitant self-report data, allowing these objective assessments to be linked to subjective states. Thus, technology has enabled a new age for collection of real-world data in real time (Kamarck et al, 1998).

5.2 Concerns About EMA

Nevertheless, there are issues that threaten the validity of these new methodologies. The frequency of EMA measurement and the fact that it takes place in participants' natural environments have raised concerns about reactivity – that is, the possibility that the act of measurement itself affects the phenomenon being measured. Evidence to date suggests that reactivity is minimal. One study randomized patients being assessed for pain to be assessed 3, 6, or 12 times daily, and it found no systematic change in their pain ratings (Stone et al, 2004), consistent with findings from an earlier study (Cruise et al, 1996). Other studies have found no effect on monitoring of behaviors such as drinking or smoking (Hufford et al, 2002). Empirical investigations have, then, reduced concern about reactivity, but further study may turn up contexts in which reactivity is a problem.

EMA studies can be demanding, often requiring participants to complete many assessments each day. This raises concerns about participants' ability or willingness to comply. Yet, across studies with diverse protocols and populations, a high degree of compliance is often achieved (Hufford and Shields, 2002). Some EMA studies make particularly high demands on participants, but what is striking is the degree of compliance observed even when the study demands might seem unrealistic on first blush. In that study where pain patients were randomized to complete 3, 6, or 12 assessments per day, compliance was excellent (averaging 94%) and was unaffected by the frequency of assessment (Stone et al, 2004). Even protocols with more than 20 prompts per day have achieved high compliance rates (Kamarck et al, 2007) Further, Freedman and colleagues (2006) showed that even homeless, crack cocaine addicts were able to complete an EMA study with multiple daily assessments with reasonable compliance. Thus, with proper management, participants seem able to bear the burden of intensive EMA sampling.

A related concern is whether the demands of EMA studies lead to bias in subject samples. We are not aware of any formal data on this, but some participants may not be willing or able to engage in these demanding protocols. In our experience, the demands of a subject's work are a common source of conflict; for example, neither surgeons nor waitresses can afford to be interrupted by unscheduled prompts. Such participant sampling bias should be evaluated and weighed in interpreting EMA data. Sometimes concerns are raised about whether older participants might have difficulty with technology such as palmtop computers. Analysis of compliance by age has demonstrated that older participants can be trained to operate the palmtops and actually demonstrate better compliance than younger participants.

There are, though, issues that may limit participants' participation. Deficits in eyesight (to see questions), hearing (to hear the phone or "beeps"), or manual dexterity (to manipulate a stylus or keypad) could certainly make some participants incapable of performing in an EMA study, though some of these deficits would also make traditional assessment difficult. More data on how EMA methods influence study participation and representativeness of subject samples would be useful.

6 Conclusion

We have argued that ecological validity is a critical component of self-report assessment for retrospective and global methods, one that is necessary for the validity of many content domains. Brunswik (1949) was correct in his assessment of the "formidable" nature of implementing representative designs to achieve what we now call "ecological validity," although he was not specifically referring to self-report data at that time. Recent developments in technology have made representative sampling of self-reports practical for most researchers, through the advent of sophisticated electronic diaries and interactive voice recording. There is no longer a need to personally shadow research participants as Brunswik did in order collect self-reports in

a representative manner to achieve ecological validity. It is our hope that knowledge of these developments will hasten the adoption of methods for collecting real-time real-world data from research participants and overcome at least some aspects of the task envisioned by Brunswik over 50 years ago.

References

Affleck, G., Tennen, H., Keefe, F. J., Lefebvre, J. C., Kashikar-Zuck, S. et al (1999). Everyday life with osteoarthritis or rheumatoid arthritis: independent effects of disease and gender on daily pain, mood and coping. *Pain, 83*, 601–609.

Blanton, H., and Jaccard, J. (2006). Arbitrary metric in psychology. *Am Psychol, 61*, 27–41.

Bolger, N., Davis, A., and Rafaeli, E. (2003). Dairy methods: capturing life as it is lived. *Ann Rev Psychol, 54*, 579–616.

Bradburn, N., Rips, L., and Shevell, S. (1987). Answering autobiographical questions: the impact of memory and inference on surveys. *Science, 236*, 151–167.

Broderick, J., Schwartz, J., and Stone, A. (2006, 3–6 May). Context (pain and affect) influences recall pain ratings [Poster presented at the Annual Meeting of the American Pain Society]. San Antonio, TX.

Broderick, J., Schwartz, J., Vikingstad, G., Pribbernow, M., Grossman, S., and Stone, A. (2008). The accuracy of pain and fatigue items across different reporting periods. Pain, *139*, 146–157.

Brunswik, E. (1944). Distal focussing of perception: size constancy in a representative sample of situations. *Psychol Monogr, 56*, 1–49.

Brunswik, E. (1949). *Systematic and Representative Design of Psychological Experiments.* Berkeley and Los Angeles: University of California Press.

Busey, T., Tunnicliff, J., Loftus, G., and Loftus, E. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychon Bull Rev, 7*, 26–48.

Byerly, M., Fisher, R., Whatley, K., Holland, R., Varghese, F. et al (2005). A comparison of electronic monitoring vs clinician rating of antipsychotic adherence in outpatients with schizophrenia. *Psychiat Res, 133*, 129–133.

Clark, D., and Teasdale, J. (1982). Diurnal variation in clinical depression and accessibility of memories of positive and negative experiences. *J Abnorm Psychol, 91*, 87–95.

Cruise, C., Porter, L., Broderick, J., Kaell, A., and Stone, A. (1996). Reactive effects of diary self-assessment in chronic pain patients. *Pain, 67*, 253–258.

Csikszentmihalyi, M., and Larsen, R. E. (1987). Validity and reliability of the experience sampling method. *J Nerv Med Dis, 175*, 526–536.

Delespaul, P. (1995). *Assessing Schizophrenia in Daily Life -- The Experience Sampling Method.* Maastricht: Maastricht University Press.

DeVries, M. (1987). Investigating mental disorders in their natural settings: introduction to the special issue. *J Nerv Men Dis, 175*, 509–513.

Eich, E., Reeves, J., Jaeger, B., and Graff-Radford, S. (1985). Memory for pain: relation between past and present pain intensity. *Pain, 223*, 375–379.

Fredrickson, B. (2000). Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions. *Cogn Emot, 14*, 577–606.

Freedman, M., Lester, K., McNamara, C., Milby, J., and Schumacher, J. (2006). Cell phones for Ecological Momentary Assessment with cocaine-addicted homeless patients in treatment. *J Subst Abuse Treat, 30*, 105–111.

Green, A., Rafaeli, E., Bolger, N., Shrout, P., and Reis, H. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychol Methods, 11*, 87–105.

Gwaltney, C., Shields, A., and Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient reported outcome measures. *Val Health , 11*, 322–333.

Hammond, K., and Stewart, T. (2001). *The Essential Brunswik: Beginnings, Explications, Applications.* New York, NY: Oxford University Press.

Hufford, M. (2007). Special methodological challenges and opportunities in Ecological Momentary Assessment. In A. Stone, S. Shiffman, A. Atienza, & L. Nebling (Eds.), *The Science of Real-Time Data Capture: Self-Reports in Health Research* (pp. 54–75). New York, NY: Oxford University Press.

Hufford, M., and Shields, A. (2002). Electronic diaries: an examination of applications and what works in the field. *Appl Clin Trials, 11*, 46–56.

Hufford, M., Shields, A., Shiffman, S., Paty, J., and Balabanis, M. (2002). Reactivity to ecological momentary assessment: an example using undergraduate problem drinkers. *Psychol Addict Behav, 16*, 205–211.

Kahneman, D., Diener, E., and Schwarz, N. (1999). *Well-Being: The Foundations of Hedonic Psychology.* New York: Russell Sage Foundation.

Kamarck, T., Shiffman, S., Smithline, L., Goodie, J., Paty, J. et al (1998). The effects of task strain, social conflict, and emotional activation on ambulatory cardiovascular activity: daily life consequences of "recurring stress" in a multiethnic sample. *Health Psychol, 17*, 17–29.

Kamarck, T., Shiffman, S., Muldoon, M., Sutton-Tyrell, K., Gwaltney, C. et al (2007). Ecological Momentary Assessment as a resource for social epidemiology. In A. Stone, S. Shiffman, A. Atienza, & L. Nebling (Eds.), *The Science of Real-Time Data Capture: Self-Reports in Health Research* (pp. 268–285). New York: Oxford University Press.

Kihlstrom, J., Eich, E., Sandbrand, D., and Tobias, B. (2000). Emotion and memory: implications for self-report. In A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman, & V. Cain (Eds.), *The Science of Self-Report: Implication for Research and Practice* (pp. 81–99). Mahwah, NJ: Erlbaum.

Korotitsch, W., and Nelson-Grey, R. (1999). An overview of self-monitoring research assessment and treatment. *Psychol Assess, 2*, 415–425.

Linton, S., and Gotestam, K. (1983). A clinical comparison of two pain scales: correlation, remembering chronic pain, and a measure of compliance. *Pain, 17*, 53–65.

Linton, S., and Melin, L. (1982). The accuracy of remembering chronic pain. *Pain, 13*, 281–285.

McFall, R. (2005). Theory and utility -- key themes in evidence-based assessment: comment on special section. *Am Psychol, 17*, 312–323.

McFarland, C., Ross, M., and DeCourville, N. (1989). Women's theories of menstruation and biases in recall of menstrual symptoms. *J Pers Soc Psychol, 57*, 522–531.

Menon, G., and Yorkston, E. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgements. In A. Stone, J. Turkkan, C. Bachrach, J. Jobe, H. Kurtzman, & V. Cain (Eds.), *The Science of Self-Report: Implications for Research and Practice* (pp. 63–79). Mahwah, NJ: Lawrence Erlbaum Associates.

Muraven, M., Collins, R., Shiffman, S., and Paty, J. (2005). Daily fluctuations in self-control demands and alcohol intake. *Psychol Addict Behav, 19*, 140–147.

Redelmeier, D., and Kahneman, D. (1996). Patients' memories of pain medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3–8.

Redelmeier, D., Katz, J., and Kahneman, D. (2003). Memories of colonoscopy: a randomized trial. *Pain, 104*, 187–194.

Robinson, M., and Clore, G. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychol Bull, 128*, 934–960.

Rock, E., Scott, J., Kennedy, D., Sridhara, R., Pazdur, R., and Burke, L. (2007). Challenges to use of health-related quality of life for Food and Drug Administration Approval of anticancer products. *J Natl Cancer Inst Monogr, 25*, 27–30.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychol Rev, 96*, 341–357.

Saito, M., Kumano, H., Yoshiuchi, K., Kokubo, N., Ohashi, K., Yamamoto, Y. et al (2005). Symptom profile of multiple chemical sensitivity in actual life. *Psychosom Med, 67*, 318–325.

Salovey, P., Sieber, W., Jobe, J., and Willis, G. (1993). The recall of physical pain. In N. Schwarz & S. Sudman (Eds.), *Autobiographical Memory and the Validity of Retrospective Reports* (pp. 89–106). New York: Springer-Verlag.

Schwarz, N. (1996). *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation.* Hillsdale, NJ: Erlbaum.

Schwarz, N. (2007). Retrospective and concurrent self-report: the rationale for real-time data capture. In A. Stone, S. Shiffman, A. Atienza, & L. Nebling (Eds.), *The Science of Real-Time Data Capture: Self-Reports in Health Research* (pp. 11–26). New York: Oxford University Press.

Schwarz, N., and Oyserman, D. (2001). Asking questions about behavior: cognition, communication and questionnaire construction. *Am J Eval, 22*, 127–160.

Schwarz, N., Wanke, M., and Bless, H. (1994). Subjective assessments and evaluations of change: some lessons learned from social cognitive research. *Eeuro Rev Soc Psychol, 5*, 181–210.

Schwarz, N., and Strack, F. (1991). Evaluating one's life: a judgment model of subjective well-being. In F. Strack, M. Argyle, & N. Schwarz (Eds.), *Subjective Well-Being: An Interdisciplinary Approach* (pp. 27–47). Oxford: Pergamon Press.

Shiffman, S. (2006). Designing protocols for Ecological Momentary Assessment. In A. Stone, S. Shiffman, A. Atienza, & L. Nebling (Eds.), *The Science of Real-Time Data Capture: Self-Reports in Health Research.* New York: Oxford University Press.

Shiffman, S. (1993). Assessing smoking patterns and motives. *J Consult Clin Psychol, 61*, 732–742.

Shiffman, S., Balabanis, M., Gwaltney, C., Paty, J., Gnys, M. et al (2007). Prediction of lapse from associations between smoking and situational antecedents assessed by ecological momentary assessment. *Drug Alc Depend, 91*, 159–168.

Shiffman, S., Hufford, M., Hickcox, M., Paty, J. A., Gnys, M., and Kassel, J. D. (1997). Remember that? A comparison of real-time vs. retrospective recall of smoking lapses. *J Consult Clin Psychol, 65*, 292–300.

Shiffman, S., Hufford, M., and Stone, A. (2008). Ecological momentary assessment in clinical psychology. *Annu Rev Clin Psychol, 4*, 1–32.

Shiffman, S., and Waters, A. (2004). Negative affect and smoking lapses: a prospective analysis. *J Consult Clin Psychol, 72*, 1192–201.

Smith, W., and Safer, M. (1993). Effects of present pain level on recall of chronic pain and medication use. *Pain, 55*, 355–361.

Stone, A., Schwartz, J., Broderick, J., and Shiffman, S. (2005). Variability of momentary pain predicts recall of weekly pain: a consequence of the peak (or salience) memory heuristic. *Person Soc Psychol Bull 31*, 1340–1346.

Stone, A., Schwartz, J., Neale, J., Shiffman, S., Marco, C., Hickcox, M. et al (1998). How accurate are current coping assessments? A comparison of momentary versus end-of-day reports of coping efforts. *J Person Soc Psychol, 74*, 1670–1680.

Stone, A., Shiffman, S., Atienza, A., and Nebling, L. (2007). *The Science of Real-Time Data Capture: Self-Reports in Health Research.* New York: Oxford University.

Stone, A., Shiffman, S., Schwartz, J., Broderick, J., and Hufford, M. (2002). Patient non-compliance with paper diaries. *Br Med J, 324*, 1193–1194.

Stone, A., Turkkan, J., Jobe, J., Bachrach, C., Kurtzman, H., and Cain, V. (2000). *The science of self report.* Mahwah, NJ: Erlbaum.

Stone, A., Broderick, J., Shiffman, S., and Schwartz, J. (2004). Understanding recall of weekly pain from a momentary assessment perspective: absolute accuracy, between- and within-person consistency, and judged change in weekly pain. *Pain, 107*, 61–69.

Stone, A. A., and Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine, 16*, 199–202.

Thompson, C., Skowronski, J., Larsen, S., and Betz, A. (1996). *Autobiographical Memory: Remembering What and Remembering When*. Mahwah, NJ: Erlbaum.

vandenBrink, M., Bandell-Hoekstra, F., and Abu-Saad, H. (2001). The occurrence of recall bias in pediatric headache: a comparison of questionnaire and diary data. *Headache, 41*, 11–20.

Wells, G., and Bradfield, A. (1998). "Good, you identified the suspect": feedback to eyewitnesses distorts their reports of the witnessing experience. *J Apply Psychol, 83*, 360–376.