Retrospective and Concurrent Self-Reports: The Rationale for Real-Time Data Capture

Norbert Schwarz
University of Michigan

Sep 2004

This chapter appeared in
A. Stone, S. S. Shiffman, A. Atienza, & L. Nebeling (eds.) (2007).
The science of real-time data capture: Self-reports in health research (pp. 11-26).
New York: Oxford University Press.

Address correspondence to Norbert Schwarz, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106-1248, USA; nschwarz@umich.edu

Self-reports are the dominant method of data collection in the social and behavioral sciences, where they are used to assess respondents' behaviors, attitudes and subjective experiences, like moods, emotions or pain. Whereas overt behaviors can, in principle, be assessed with other methods, individuals' self-reports provide the only window on their inner states. Unfortunately, this window is often foggy, in particular when respondents are expected to report on extended time periods. For example, the Health Interview Survey conducted by the National Center for Health Statistics asks respondents, *"Now, I'd like to read you a short list of different kinds of pain. Please say for each one, on roughly how many days -- if any -- in the last 12 months you have had that type of pain. How many days in the last year have you had headaches? [Question repeated for backaches; stomach pains; joint pains; muscle pains; dental pains]"* I encourage readers to answer this question. Trying to do so quickly raises another question: Are we asking people for things that they can't tell us?

This chapter addresses what people can and can not tell us and reviews what we know about the cognitive and communicative processes underlying retrospective self-reports. It draws on research at the interface of autobiographical memory, judgment and research methodology and highlights the numerous sources of bias that threaten the validity of retrospective reports (for comprehensive reviews see Schwarz & Sudman, 1994; Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). Some of these biases can be attenuated through the use of better interviewing techniques, like Event History Calendars (see Belli, 1998), but the room for improvement is limited by the limits of autobiographical memory. As an alternative, we can ask people to report on things they *can* tell us about, namely their *current* behavior and experiences. This is the promise of real-time data capture, which focuses on the collection of concurrent rather than retrospective self-reports. Real-time data capture, exemplified by ecological momentary assessment (EMA; see Stone, Shiffman, & DeVries, 1999), attenuates or eliminates the memory related biases inherent in retrospective reports. Unfortunately, it does not eliminate other potential sources of bias in self-reports, from problems of question comprehension to the influence of question order or response alternatives (for reviews see Schwarz, 1999a; Sudman et al., 1996).

Table 1

Table 1 shows the main types of retrospective self-reports and summarizes key points discussed in the following sections. I first review the cognitive and communicative processes underlying these reports and identify what respondents can and can not tell us. Throughout, I highlight how the underlying processes result in systematic bias. Following this review of retrospective self-reports, I turn to the cognitive and communicative processes involved in concurrent self-reports and identify open issues for future research.

Historical Information

Some retrospective questions pertain to historical information. Examples include, *Have you ever had an episode of back pain? In what year did you first have an episode of back pain?* Respondents' memories are usually the only available source of information on these issues and real-time data capture does not provide an alternative strategy. The best a researcher can do is to use interviewing techniques that take the structure of autobiographical memory into account to facilitate recall (for advice see Belli, 1998; Schwarz & Oyserman, 2001; Tourangeau, et al., 2000).

The structure of autobiographical memory can be thought of as a hierarchical network that includes *extended periods* (like "the years I lived in New York") at the highest level of the hierarchy. Nested within this high-order period are lower-level extended events pertaining to this time, like "my first job" or "the time I was married to Lucy." Further down the hierarchy are *summarized events*, which correspond to the knowledge-like representations of repeated behaviors noted above (e.g., "During that time, I was frequently ill."). *Specific events*, like a particular episode of illness, are represented at the lowest level of the hierarchy. To be represented at this level of specificity, however, the event has to be rather unique. As these examples illustrate, autobiographical memory is primarily organized by time ("the years in New York") and relatively global themes ("first job;" "first marriage;" "illness") in a hierarchical network (see Belli, 1998, for a comprehensive review). This network "permits the retrieval of past events through multiple pathways that work top-down in the hierarchy, sequentially within life themes that unify extended events, and in parallel across life themes that involve contemporaneous and sequential events" (Belli, 1998, p. 383). Such searches take

considerable time and their outcome is somewhat haphazard, depending on the entry point into the network at which the search started. Hence, using multiple entry points and forming connections across different periods and themes improves recall.

These fortunate conditions are rarely met in research interviews. One promising method to instantiate them is offered by the *event history calendar* (see Belli, 1998, for a comprehensive review). It allows respondents to place their behavior in time and space and uses the hierarchically nested structure of autobiographical memory to facilitate recall. Moreover, it provides respondents with considerable time for the recall task and emphasizes the importance of accuracy. Finally, it explicitly encourages the correction of earlier answers as newly recalled information qualifies earlier responses. This correction opportunity is missed under regular interview formats, where respondents can rarely return to earlier questions.

To assess a respondent's health-history, for example, respondents may begin by marking life-periods like being in school, living at home, getting a first job, and so on. Next, they may be asked to mark health-relevant events within these periods, changing the timing of already marked events as needed when newly recalled information requires corrections. Within the developing rich structure of associations, respondents are usually able to recall and date events with considerable accuracy (e.g., Freedman et al., 1996; Caspi et al., 1996). Although costly in terms of interview time, this approach provides the most promising avenue for the assessment of historical information. Without such efforts, recall errors are highly likely as illustrated by Cannell, Fisher, and Bakker's (1965) observation that 42% of a sample of patients failed to report an episode of overnight hospitalization when interviewed one year after the event.

Frequency Reports

Frequency questions ask respondents to report on the frequency of a behavior or experience during a specified reference period, often last week or last month. Researchers typically hope that respondents will identify the behavior of interest, search the reference period, retrieve all instances that match the target behavior, and finally count these instances to determine the overall frequency of the behavior. However, respondents can only follow such a *recall-and-count* strategy under very limited

circumstances. In most cases, they need to rely on extensive inference and estimation strategies to arrive at an answer. Which strategy they use depends on the frequency, importance, and regularity of the behavior (e.g., Brown, 2002; Menon, 1993, 1994; Sudman et al., 1996).

Rare and Important Behaviors

Rare and important behaviors can be reported on the basis of autobiographical knowledge or a recall-and-count strategy. When asked "*How often did you get divorced?"* most people can provide an accurate answer without extended memory search. On the other hand, when asked *"How often did you relocate to another city?"* many respondents do not know immediately but can determine the appropriate answer by reviewing their educational and job history, following a recall-and-count strategy. Real-time data capture is not suited for such tasks, due to the low frequency of the respective behavior and the extended time periods covered. Fortunately, respondents' answers will nevertheless be relatively accurate when the behavior is rare and important, provided that they are encouraged to take the time necessary for extensive recall.

Frequent Behaviors

Respondents' task is considerably more demanding when the behavior is frequent. In this case, respondents are unlikely to have detailed representations of numerous individual episodes of a behavior stored in memory. Instead, the various instances of closely related behaviors blend into one global, knowledge-like representation that lacks specific time or location markers (see Linton, 1982; Strube, 1987). Frequent doctor visits, for example, result in a well-developed knowledge structure for the general event, allowing respondents to report in considerable detail on what usually goes on during their doctor visits. But the highly similar individual episodes become indistinguishable and irretrievable, making it difficult to report on any specific one. This is most likely to occur for mundane behaviors of high frequency, but has also been observed for more important events. Mathiowetz and Duncan (1988), for example, found that respondents were more accurate in recalling a single spell of unemployment than they were at recalling multiple spells of unemployment.

As a result of the knowledge-like representation of frequent behaviors and experiences, respondents cannot rely on a recall-and-count strategy. Instead, they need to

resort to estimation strategies to arrive at a plausible frequency report. Which strategy they choose depends on the regularity of the behavior and the context in which the frequency question is presented.

Rate Information and Extrapolation. When the behavior is highly regular, respondents can arrive at a frequency estimate on the basis of rate information (Menon, 1994; Menon, Raghubir, & Schwarz, 1995). Respondents who go to church every Sunday, or wash their hair every day, face little difficulty in computing a weekly or monthly estimate. Unfortunately, exceptions are likely to be missed and the obtained answers are only accurate when the behavior does indeed conform to the assumed rate. A related estimation strategy relies on extrapolation from partial recall. When asked how often she took pain medication during the last week, for example, a respondent may reason, "I took pain killers three times today, but this was a bad day. So probably twice a day, times 7 days, makes 14 times a week." The accuracy of this estimate will depend on the accuracy of the underlying assumptions, the regularity of the behavior, and the specifics of the day that served as input into the chain of inferences.

As these examples illustrate, respondents are unlikely to answer frequency questions on the basis of the recall-and-count strategy that most researchers hope for. Instead, their answers are, at best, based on partial recall and extensive inferences, unless the behavior is rare and important, and hence highly memorable (for extended discussions see Brown, 2002; Schwarz & Oyserman, 2002). Other estimation strategies may even bypass any effort to recall specific episodes. One such strategy is respondents' reliance on information provided by the research instrument itself.

Frequency Scales. Respondents are often asked to report the frequency of their behavior by checking the appropriate alternative from a list of quantitative response alternatives of the type shown in Table 2. What is typically overlooked is that respondents believe that researchers construct meaningful scales that are relevant to the task at hand. Specifically, they assume that the scale reflects the researcher's knowledge about the distribution of the behavior, with values in the middle range of the scale corresponding to the "usual" or "average" behavior and values at the extremes of the scale corresponding to the extremes of the distribution. Given these assumptions, respondents can use the range of the response alternatives as a frame of reference in estimating their own behavioral

frequency. This results in higher frequency estimates along scales that present high rather than low frequency response alternatives.

For example, Schwarz and Scheuring (1992) asked 60 patients of a German mental health clinic to report the frequency of 17 symptoms along one of the two scales shown in Table 2. Across 17 symptoms, 62% of the respondents reported average frequencies of more than twice a month when presented with the high frequency scale, whereas only 39% did so when presented with the low frequency scale, resulting in a mean difference of 23 percentage points. The impact of response alternatives was strongest for the ill-defined symptom of "responsiveness to changes in the weather," where 75% of the patients reported a frequency of more than twice a month along the high frequency scale, whereas only 21% did so along the low frequency scale. Conversely, the influence of response alternatives was least pronounced for the better defined symptom "excessive perspiration," with 50% vs. 42% of the respondents reporting a frequency of more than twice a month in the high and low frequency scale conditions, respectively. The differential size of the observed scale effects also illustrates that respondents are more likely to resort to estimation strategies, the less salient and memorable the respective behavior is (e.g., Schwarz, 1999).

Higher frequency reports along high rather than low frequency scales have been observed across a wide range of behaviors, including health behaviors (e.g., Gaskell, O'Muircheartaigh, & Wright, 1994), television consumption (e.g., Schwarz, Hippler, Deutsch, & Strack, 1985), sexual behaviors (e.g., Tourangeau & Smith, 1996), and consumer behaviors (e.g., Menon, et al., 1995). Several methodological implications deserve attention: First, these findings call the meaning of the absolute reports into question and show that reports provided along different scales are not comparable. Second, they illustrate that the impact of frequency scales is more pronounced the more poorly the behavior is represented in memory, as expected on theoretical grounds (Menon et al., 1995). When behaviors of differential memorability are assessed, this can either exaggerate or cloud actual differences in the relative frequency of the behaviors, undermining comparisons across behaviors. Third, respondents with poorer memory are more likely to be influenced than respondents with better memory (e.g., Knäuper, Schwarz, & Park, 2004; Schwarz, 1999b). This can undermine comparisons across groups that differ

in memory performance (e.g., younger vs. older respondents; see Knäuper et al., 2004). Finally, for any given group, the use of frequency scales results in an underestimation of the variance in behavioral frequencies. Because all respondents draw on the same frame of reference in computing an estimate, the estimates they compute are more similar than warranted.

Summary

As this selective review illustrates, retrospective frequency reports are fraught with uncertainty. In particular when the behavior is frequent, mundane, and irregular, respondents can only arrive at a frequency report by relying on an estimation strategy. Unfortunately, many of the behaviors that health researchers are interested in fit these characteristics. Under these conditions, concurrent reports are highly preferable.

Intensity Reports

Other retrospective questions pertain to the intensity of an experience (*How pleasant, painful, etc. was it?).* Intensity reports show pronounced biases even after a very short delay, making real-time data capture the method of choice. In general, our subjective experiences, including the intensity of our feelings, are poorly represented in memory: Once the experience ends, its characteristics can no longer be directly examined (see Robinson & Clore, 2002, for a review). Accordingly, respondents again need to rely on fragmentary recall and extensive inferences to arrive at an answer. I first address which moments of an extended experience are particularly likely to be remembered and how these moments influence the overall retrospective recall. Subsequently, I turn to respondents' inference strategies.

Memorable Moments: Peak & End Effects

Redelmeier and Kahneman (1996) asked patients undergoing a colonoscopy to provide concurrent ratings of their pain along a 10-point scale. Figure 1 shows the pain ratings of two patients, collected at 60 second intervals. Both patients experienced similar peak pain, but patient B's colonoscopy lasted more than twice as long than patient A's colonoscopy, resulting in an overall experience of more pain. A few minutes after the completion of the procedure, however, patient B evaluated the overall experience as *less* painful.

Figure 1

As numerous related studies demonstrated (for a review see Fredrickson, 2000), this surprising result reflects that retrospective assessments follow a *peak-and-end* heuristic. This heuristic draws on two pieces of information that are particularly relevant from a survival point of view, namely the peak ("How bad does it get?") and end ("How does it end?"). Hence, patient B was left with a memory of "less pain" than patient A: Although both experienced the same peak pain, the final moment of patient B's colonoscopy was more benign, resulting in a more favorable memory despite similar peak pain and longer pain duration.

Findings of this type have two important implications. First, they illustrate that the duration of experiences is not well represented in memory and hence largely neglected in retrospective reports, even after a very short delay. Second, they highlight that retrospective reports of the intensity of an experience are based on two distinct moments, the peak and end, while neglecting the intensity at other moments. Quite clearly, the concurrent measurement of pain, shown in Figure 1, provides a more accurate picture of these patients' experience than their retrospective reports. Reiterating this theme, Stone and colleagues observed peak-and-end effects in the retrospective pain reports of arthritis patients over a one-week period (Stone, Broderick, Porter, & Kaell, 1997).

Despite their profound advantages in capturing people's actual experiences, concurrent measures may not always be the measure of choice. Suppose, for example, that the goal is to predict patients' compliance with the need for a subsequent follow-up colonoscopy. In light of his more benign memory of the procedure, patient B is more likely to comply than patient A, in contrast to what the concurrent measures of pain may suggest (for examples see Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Redelmeier, Katz, & Kahneman, 2003). Given that people's decisions are based on their *remembered* experiences, retrospective reports may often be better predictors of their behavior, even though concurrent reports excel at capturing what was really going on. From a methodological perspective, this implies that a close link between retrospective reports and behavioral intentions does not validate the retrospective report; nor does a poor relationship between concurrent reports and behavioral intentions invalidate the

concurrent reports. Instead, both observations merely reflect that we don't "learn from experience" – we learn from remembered experience.

Inference Strategies: The Role of Naïve Theories

In most cases, respondents have to rely on inference strategies to arrive at retrospective reports of intensity. Suppose a respondent is asked how bad her pain was last week. To answer this question, the respondent may use her present pain as benchmark, asking herself if there is any reason to assume that last week was different. If she recalls such a reason, she will adjust her retrospective report accordingly; if not, she will report her current pain as a good approximation of last week's pain. As a result, her retrospective report of pain is a function of her current pain and her naïve theory about the stability of her pain over time.

In many domains, individuals assume an unrealistically high degree of stability, resulting in retrospective reports about past behaviors and experiences that are more similar to their present behaviors or experiences than is warranted. Accordingly, retrospective estimates of income (Withey, 1954) and of tobacco, marijuana, and alcohol consumption (Collins, Graham, Hansen, & Johnson, 1985) were found to be heavily influenced by respondents' income or consumption habits at the time of interview. The same holds for retrospective reports of pain. For example, Eich and colleagues (1985) asked chronic pain patients to report their current pain and the maximum, minimum, and usual pain of they experienced during the previous week. When they compared these retrospective reports with patients' concurrent entries in daily pain diaries, they observed a familiar pattern: The retrospective reports were more similar to patients' pain at the time of recall than warranted. Thus, high current pain resulted in an overestimation of last week's pain, whereas low current pain resulted in an underestimation.

On the other hand, when respondents have reason to believe in change, they will detect change, even though none has occurred (see Ross, 1989). For example, Ross and Conway (1986) had students participate in a study skills training that did not improve their skills on any objective measure (and was not expected to do so). Following the training, researchers asked participants to recall how skilled they were before the training. Applying a plausible theory of change, namely that the training improved their skills, participants inferred that their prior skills must have been much worse than they were after training.

Hence, they retrospectively reported having had poorer pre-training skills than they indicated before the training, apparently confirming the intervention's success. This result was obtained despite incentives to respondents to recall their earlier answers as accurately as possible. As Ross and Conway (1986) noted, you can always get what you want by revising what you had. The same logic, again, applies to retrospective reports of pain intensity. For example, Linton and Melin (1982) had back pain patients record their pain prior to treatment program (baseline measurement). Following program completion, the patients were asked to recall their baseline pain. Reiterating Ross and Conway's (1986) finding, they now "recalled" more baseline pain than they had reported concurrently, apparently confirming the success of the treatment program. Again, concurrent measures of pain would provide a more accurate assessment.

Summary

The intensity of experiences is poorly represented in memory. Accordingly, respondents resort to inference strategies unless the episode is highly memorable. Even in the latter case, however, their retrospective reports are likely to be based on only a few moments of the episode, namely its peak and end. Accordingly, concurrent measures are preferable whenever their collection is feasible.

Reports of Change

The tasks posed to respondents sometimes go beyond mere retrospective reports. This is the case when researchers attempt to compensate for the lack of longitudinal data by asking respondents to report on change over time: *Do you have more or less pain now than you had at the beginning of the treatment?* As already seen in the preceding section, respondents' theory-driven inferences can make the past seem more or less similar to the present than warranted, resulting in systematic biases. Such theory-driven inferences are particularly likely, and problematic, when the context suggests an applicable theory, as is often the case in medical studies: Believing that things get better with treatment (or why else would one undergo it?), patients are likely to infer that their past condition must have been worse than their present condition (e.g., Linton & Melin, 1982; for a review see Ross, 1989). This reliably results in overly optimistic assessments of change, which may explain the recent popularity of subjective change reports as "patient reported outcomes."

From a cognitive perspective, asking patients whether they feel better now than before their treatment is the most efficient way to "improve" the success rate of medical interventions.

As this discussion indicates, there is no substitute for appropriate study design and respondents can not compensate for the researchers' earlier oversights or lack of funds. If change over time is of crucial interest, concurrent measures at different points in time are the only reliable way to assess it.

Reports of Covariation and Causation

Similar problems arise when respondents are asked to report on covariation (*Under which circumstances…?*) or causation (*Why…?*). To arrive at an observation based answer to these questions, respondents would need to have an accurate representation of the frequency of their behaviors, the different contexts of these behaviors, and the intensity of related experiences. As already seen, respondents are often unable to provide accurate reports on any of these components, making their joint consideration an unrealistically complex task. Once again, respondents are likely to draw on applicable naïve theories, resulting in systematic biases.

As an example, consider a study by McFarland and colleagues (1989), in which women kept daily diaries of their affect and physical symptoms. Following the diary period, the women were asked to recall their affect and physical symptoms for a particular day, either a day during their menstrual or their intermenstrual phase. A comparison of their concurrent diary data and their retrospective reports showed a pronounced impact of their beliefs about menstruation. Women who considered their menstruation a distressing event recalled feeling worse during the menstrual phase than they had reported concurrently, whereas these beliefs did not bias recall for a day during the intermenstrual phase.

Findings of this type highlight that the reconstructed experience is itself a function of naïve theories of covariation ("I feel bad during menstruation"). This reconstruction, in turn, "confirms" the naïve theory, rendering its future application more likely. Note that

this process is the opposite of what researchers hope for: Instead of inferring covariation from accurately recalled behaviors and circumstances, respondents often draw on naïve theories of covariation to reconstruct the relevant behaviors in the first place. Accordingly, reports of covariation and causation are fraught with uncertainty, unless they pertain to relatively simple and highly memorable events ("I broke my leg because I fell off the ladder").

Again, there is no substitute for appropriate study design and covariation and causation are best assessed with real-time data capture. EMA excels at this task by prompting respondents to report on their behavior, experiences, and circumstances, allowing researchers to collect all the data needed for appropriate analyses, as the contributions to this volume illustrate. As already noted in the context of intensity reports, however, this does not render respondents' beliefs about covariation irrelevant. Respondents' behavioral decisions are likely to be driven by their own beliefs about covariation, inaccurate as those beliefs may be. Once again, assessing reality and capturing people's perceptions of reality are different enterprises, with different purposes and pay-offs.

Real-Time Data Capture: Promises and Challenges

As this selective review indicates, retrospective questions often ask respondents for information that they cannot provide with any validity. The key promise of real-time data capture methods is that they pose more realistic tasks by asking respondents for information they know: their current behavior, experiences, and circumstances. However, respondents' memory limitations are not the only source of biases in self-reports and issues of question comprehension, response formatting or social desirability arise in concurrent as well as retrospective reports (for reviews see Schwarz, 1999a; Sudman et al., 1996). To date, these complications have received little attention in methodological studies of real-time data capture. In this final section, I illustrate the needed research with three examples.

Question Comprehension

To provide a meaningful answer, respondents typically have to go beyond the literal meaning of a question to infer what the researcher is interested in. To do so, they make extensive use of contextual information, including features of the question that the researcher may consider tangential to the question's meaning (for reviews see Clark & Schober, 1992; Schwarz, 1996). Suppose, for example, that a respondent is asked how often he has been "angry" recently. To answer this question, the respondent needs to determine what kind of "anger" the researcher has in mind: minor irritations or major annoyances? One source of information that respondents draw on is the reference period specified in the question (Winkielman, Knäuper, & Schwarz, 1998). When the question pertains to "yesterday," respondents assume that the researcher has minor irritations in mind because "big anger" doesn't happen that often. Conversely, when the question pertains to the "last six months," respondents assume that the researcher has major annoyances in mind because they can hardly be expected to recall all the minor irritations of life over such an extended period. As a result, respondents deliberately report on substantively different experiences, depending on the reference period specified in an otherwise identical question (Winkielman et al., 1998; Igou, Bless, & Schwarz, 2000).

This observation has potentially profound implications for real-time data capture, where the typical reference period is very short ("right now" or the last few hours). It suggests that EMA respondents may include relatively minor events in their reports, which would go unmentioned for any longer reference period. This renders differences between concurrent and retrospective frequency reports of "anger," for example, highly ambiguous. On the one hand, concurrent reports may indicate higher frequencies than retrospective reports because some episodes were simply forgotten. On the other hand, concurrent reports may do so because they include many minor episodes that respondents may consider irrelevant when asked a retrospective question, pertaining to a longer reference period. In the latter case, concurrent and retrospective reports would pertain to subjectively "different" questions, rendering the answers noncomparable.

Scale Use

A related issue arises with regard to scale use. When asked to indicate how angry they are along a rating scale, respondents need to determine the meaning of the scale numbers. They do so by anchoring the endpoints of the scale with low and high anger events. Accordingly, their current anger receives a lower rating, the more extreme the event is that serves as the high anchor (e.g., Parducci, 1965; Daamen & de Bie, 1992). Retrospective and concurrent assessments are likely to differ in terms of the comparison episodes that come to mind. When asked to rate a single past episode, the recalled episode is likely to be compared to other memorable instances – which are often memorable because they were extreme. But when asked to rate multiple episodes over the course of a single day, previously rated moderate episodes may still be highly accessible. Accordingly, retrospective and concurrent ratings may differ in the comparison points and scale anchors used, undermining the comparability of the obtained results.

Social Desirability

Not surprisingly, respondents sometimes hesitate to report behaviors or opinions that they consider undesirable (for a review see DeMaio, 1984). It seems likely, however, that such social desirability concerns are less pronounced for reports that pertain to very short reference periods. For example, a parent may well report *"I couldn't stand my kids last night,"* a report that pertains to a specific, limited episode. Yet the same parent may hesitate to report *"I don't like being with my kids,"* given the more pervasive implications of this general statement. If so, it is conceivable that concurrent reports, pertaining to limited episodes, are less subject to social desirability bias and more likely to capture negative and undesirable thoughts and feelings. But how many episodic reports does it take until respondents become aware that reporting that they don't like being with their kids "right now" amounts to the same as reporting that they rarely like being with them? Does socially desirable responding increase over repeated measurements?

Research Needs

At present, we know nothing about the conjectures offered above. To date, methodological research on real-time data capture has largely focused on overcoming the

limits of retrospective reports. While this focus is important, we also need systematic experimental research into the cognitive and communicative processes underlying *concurrent* reports. Without this work, we run the risk of merely replacing the known biases of retrospective reports with new, unknown biases of concurrent reports.

Concluding Remarks

As this review illustrates, psychologists and survey methodologists have made considerable progress in understanding the cognitive and communicative processes underlying self-reports (for more comprehensive treatments see Schwarz, 1999a; Sudman et al., 1996; Tourangeau et al., 2000). Despite the progress made, however, human memory imposes limits on what people can validly report on. Under most of the conditions of interest to health researchers, respondents have to rely on partial recall and extensive inference strategies when asked to report on their past behavior and experiences. These strategies result in biases that are well understood and difficult to avoid. Methods of real-time data capture provide a promising alternative by asking people about things that they *can* report on: their current behavior and current experiences. Moreover, EMA as the prime example of real-time data capture methods offers unique opportunities to assess behaviors and experiences in their ecological context and to complement subjective reports with objective measures, as the contributions to the present volume illustrate. To fully develop the potential of real-time data capture, however, future research will need to attend to the cognitive and communicative processes underlying concurrent reports, thus complementing previous work on retrospective reports.

References

Belli, R. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys, Memory, 6, 383-406.

Brown, N. R. (2002). Encoding, representing, and estimating event frequencies: Multiple strategy perspective. In P. Sedlmeier & T. Betsch (Eds.), Frequency processing and cognition (pp. 37-54) New York: Oxford University Press.

Cannell, C. F., Fisher, G., & Bakker, T. (1965). Reporting on hospitalization in the Health Interview Survey. Vital and Health Statistics (PHS Publication No. 1000, Series 2, No. 6). Washington, D.C.: US Government Printing Office.

Caspi, A., Moffitt, T., Thornton, A., Freedman, D., Amell, J., Harrington, H., Smeijers, J. & Silva, P. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. International Journal of Methods in Psychiatric Research, 6, 101-114.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), Questions about questions (pp. 15-48). New York: Russel Sage.

Collins, L.M., Graham, J.W., Hansen, W.B., & Johnson, C.A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. Applied Psychological Measurement, 9, 301 - 309.

Daamen, D. D.,L., & de Bie, S. E. (1992). Serial context effects in survey items. In N. Schwarz & S. Sudman (Eds.), Context effects in social and psychological research (pp. 97-114). New York: Springer Verlag.

DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), Surveying subjective phenomena (Vol. 2, pp. 257-281). New York: Russell Sage.

Eich,E., Reeves, J.L., Jaeger, B., & Graff-Radford, S.B. (1985). Memory for pain: Relation between past and present pain intensity. Pain, 23, 375-380.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. In C.C. Clogg (Ed.) Sociological Methodology, vol 18, (pp. 37-68). Washington, D.C.: American Sociological Association.

Fredrickson, B.L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. Cognition & Emotion, 14, 577-606.

Gaskell, G. D., O'Muircheartaigh, C. A., & Wright, D. B. (1994). Survey questions about the frequency of vaguely defined events: The effects of response alternatives. Public Opinion Quarterly, 58, 241-254.

Igou, E.R., Bless, H., & Schwarz, N. (2002). Making sense of standardized survey questions: The influence of reference periods and their repetition. Communication Monographs, 69, 179-187.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A., & Redelmeier, D. (1993). When more pain is preferred to less: Adding a better end. Psychological Science, 4, 401-405.

Knäuper, B., Schwarz, N., & Park, D.C. (2004). Frequency reports across age groups: Differential effects of frequency scales. Journal of Official Statistics, 20, 91-96.

Linton, M. (1982). Transformations of memory in everyday life. In U. Neisser (Ed.), Memory observed: Remembering in natural contexts (pp.77-91). San Francisco: Freeman.

Linton, S. J., & Melin, L. (1982). The accuracy of remembering chronic pain. Pain, 13, 281-285.

Mathiowetz, N.A., & Duncan, G.J. (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. Journal of Business and Economic Statistics, 6, 221-229.

McFarland, C., Ross, M., & De Courville (1989). Women's theories of menstruation and biases in recall of menstrual symptoms. Journal of Personality and Social Psychology, 57, 522-531.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. Journal of Consumer Research, 20, 431-440.

Menon, G. (1994). Judgments of behavioral frequencies: Memory search and retrieval strategies. In N. Schwarz & S. Sudman, S. (Eds.) (1994). Autobiographical memory and the validity of retrospective reports (pp. 161- 172). New York: Springer Verlag.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. Journal of Consumer Research, 22, 212-228.

Parducci, A. (1965). Category judgment: A range-frequency model. Psychological Review, 72, 407-418.

Redelmeier, D., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. Pain, 116, 3-8.

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

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. Psychological Bulletin, 128, 934-960.

Ross, M. (1989). The relation of implicit theories to the construction of personal histories. Psychological Review, 96, 341-357.

Ross, M., & Conway,, M. (1986). Remembering one's own past: The construction of personal histories. In R. M. Sorrentino & E.T. Higgins (Eds.), Handbook of motivation and cognition (pp. 122 - 144). New York: Guilford.

Schwarz, N. (1996). Cognition and communication: Judgmental biases, research methods, and the logic of conversation. Hillsdale, NJ: Erlbaum.

Schwarz, N. (1999a). Self-reports: How the questions shape the answers. American Psychologist, 54, 93-105.

Schwarz, N. (1999b). Frequency reports of physical symptoms and health behaviors: How the questionnaire determines the results. In Park, D.C., Morrell, R.W., & Shifren, K. (Eds.), Processing medical information in aging patients: Cognitive and human factors perspectives (pp. 00-00). Mahaw, NJ: Erlbaum.

Schwarz, N., Hippler, H.J., Deutsch, B. & Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. Public Opinion Quarterly, 49, 388-395.

Schwarz, N. & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication and questionnaire construction. American Journal of Evaluation, 22, 127-160.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Anwortvorgaben des Fragebogens lernen. (Frequency-reports of psychosomatic symptoms: What respondents learn from response alternatives.) Zeitschrift für Klinische Psychologie, 22, 197-208.

Schwarz, N. & Sudman, S. (1994). Autobiographical memory and the validity of retrospective reports. New York: Springer Verlag.

Schwarz, N., & Sudman, S. (1996). Answering questions: Methodology for determining cognitive and communicative processes in survey research. San Francisco: Jossey-Bass.

Stone, A. A., Broderick, J. B., Porter, L., & Kaell, A. T. (1997). The experience of rheumatoid arthritis pain and fatigue: Examining momentary reports and correlates over one week. Arthritis Care and Research, 10, 185-193.

Stone, A.A., Shiffman, S. S., & DeVries, M. W. (1999). Ecological momentary assessment. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), Well-being: The foundations of hedonic psychology (pp. 61-84). New York: Russell-Sage.

Strube, G. (1987). Answering survey questions: The role of memory. In H.J. Hippler, N. Schwarz, & S. Sudman (Eds.), Social information processing and survey methodology (pp. 86 - 101). New York: Springer Verlag.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, CA: Jossey-Bass.

Tourangeau, R., Rips, L.J., & Rasinski,K. (2000). The psychology of survey response. New York: Cambridge University Press.

Tourangeau, R., & Smith, T.W. (1996). Asking sensitive questions. The impact of data collection, mode, question format, and question context. Public Opinion Quarterly, 60, 275-304.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of (emotion) frequency questions. Journal of Personality and Social Psychology.

Withey, S. B. (1954). Reliability of recall of income. Public Opinion Quarterly, 18, 31 - 34.

Table 1. Types of Information Sought

Type	**Real-time Data Capture**	**Retrospective Reports**
Historical Information *Ever? First?*	Not applicable	Feasible; reporting can be improved through the methodology of event history calendars
Frequency *How often?*		
a. Rare	Not applicable	Feasible when behaviors are distinct and important
b. Frequent and regular	Applicable but not needed	Extrapolation from rate information feasible
c. Frequent and irregular	Applicable and preferable	Estimation strategies dominate; bias likely.
Intensity *How intense, pleasant, painful, etc.?*	Applicable and preferable	Likely to be biased, even after a short delay
Change over time *More or less…?*	Applicable and preferable, provided behavior is frequent	Theory-driven and biased
Covariation/causation *When and why?*	Applicable and preferable, provided behavior is frequent	Theory-driven and biased

Table 2. Frequency Response Alternatives for Reporting Physical Symptoms

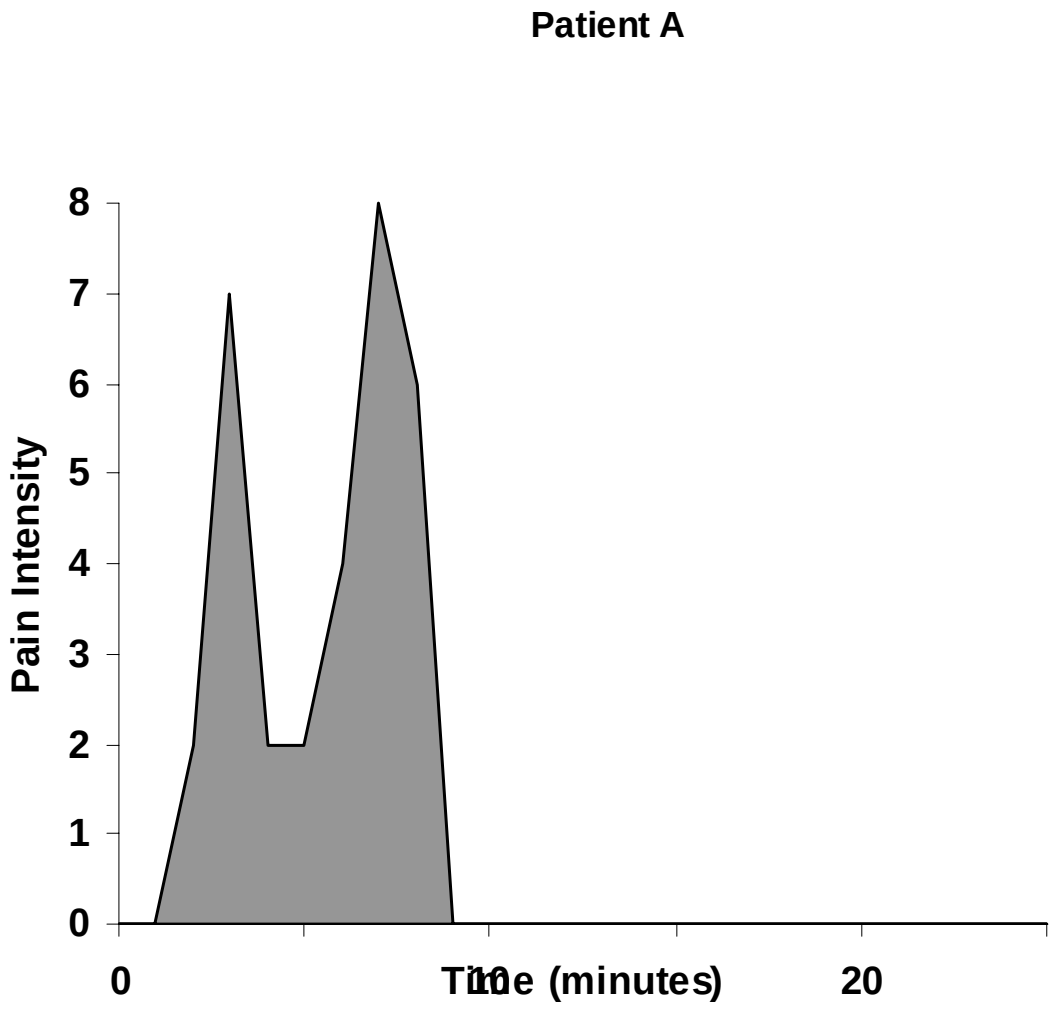
Low Frequency Scale	**High Frequency Scale**
( ) never	( ) twice a month or less
( ) about once a year	( ) once a week
( ) about twice a year	( ) twice a week
( ) twice a month	( ) daily
( ) more than twice a month	( ) several times a day

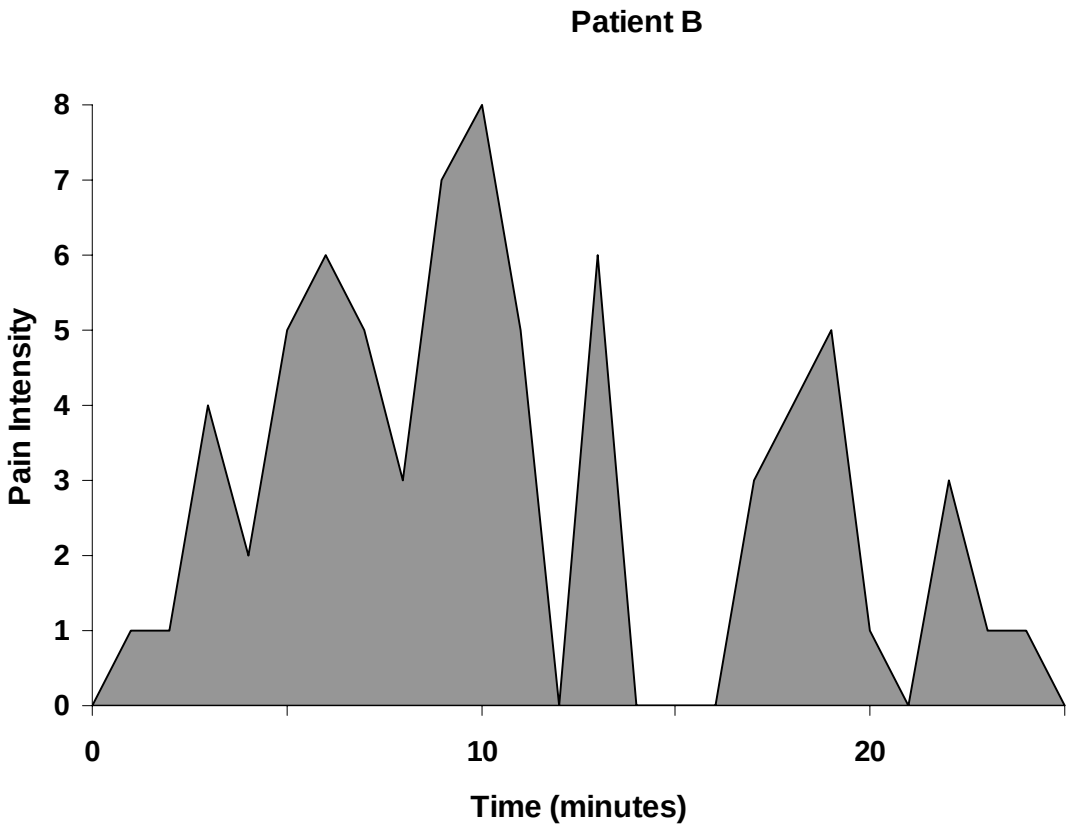
Figure Captions

Figure 1: Concurrent Reports of Pain

Note: Shown are real-time pain reports from two patients undergoing colonoscopy. The x-axis represents time in minutes from the start of the procedure; the y-axis represents the intensity of pain reported in real-time on a visual analog scale (0 = no pain; 10 = extreme pain). Adapted from Kahneman & Redlmeier, 1996. Reprinted by permission.

To Editors:
This figure consists of 2 panels, "Patient A" and "Patient B." Both panels are attached as pdf files. The graphs were provided by the original author (Kahneman); permission of the publisher has been requested.


Patient A
Pain Intensity
8
7
6
5
4
3
2
1
0
0
10
20
Time (minutes)


Patient B
Pain Intensity
0
1
2
3
4
5
6
7
8
0
10
20
Time (minutes)