The Health and Retirement Study - An Evaluation and Scientific Ideas for the Future

Time Use and Well-being, and Large Survey Studies

Dylan M. Smith*

*Stony Brook University, dylan.m.smith@stonybrook.edu
Time Use and Well-being, and Large Survey Studies*

Dylan M. Smith

Abstract

This paper reviews several methods for measuring how people spend their time, and how they feel during these different activities, and argues that some of these methods could be well suited for large scale longitudinal surveys. Because time use methods allow for the quantitative assessment of the dynamics of human experience, they provide opportunities to explore numerous research questions that cannot be readily answered with more traditional summary measures of well-being. In the last decade or two, techniques have become available that can capture the dynamics of time use and well-being in ways that also reduce methodological problems such as biased recall. The paper describes several such methods, with a discussion of how they are implemented, and a comparison of their relative strengths and weaknesses. In the final section, the paper describes how the addition of time use measures can enhance national surveys such as the Health and Retirement Study by a) opening new avenues of research, and b) clarifying previous findings from more traditional measures of well being.

KEYWORDS: Health and Retirement Study

*Dylan M. Smith is Associate Professor at the Center for Medical Humanities, Compassionate Care, and Bioethics, School of Medicine, Stony Brook University. Please send correspondence to dylan.m.smith@stonybrook.edu. The author would to thank Arthur Stone and Stephanie Brown for providing editorial feedback on this paper, feedback which resulted in substantial improvements.
Introduction

In 2004, the Health and Retirement Study (HRS) substantially expanded its utilization of psychosocial measures, adding a number of important variables such as social support, life satisfaction, optimism, and purpose in life, among others, administered after interviews in the form of a ‘leave-behind’ survey. Prior to this expansion, the inclusion of psychosocial measures was modest (Ryff 2002), especially in the area of well being. Indeed, there was no direct measure of well being (though some investigations created an ad hoc measure by combining several items from the CES-d measure of depression (e.g., Smith, Langa et al. 2005). The recent additions greatly improve the ability of investigators to measure various aspects of well being in this sample, creating multiple new avenues for research (Ryff, 2002).

In this brief review, I will describe several methods for capturing time use; that is, measures that assess how people spend their time, and how they feel during these different activities, and argue that the ability of the HRS and other large scale surveys to capture important dimensions of well being would be further enhanced by the inclusion of some form of time use measure. Because time use methods allow for the quantitative assessment of the dynamics of human experience, they provide opportunities to explore numerous research questions that cannot be readily answered with more traditional summary measures of well being. In the last several years, new techniques have become available that can capture the dynamics of time use and well being in ways that also reduce methodological problems such as biased recall. Some of the techniques reviewed here present high levels of respondent (and researcher) burden, but one of the most recently developed techniques—the Day Reconstruction Method (Kahneman, Krueger et al. 2004)—has relatively low burden and has already been successfully fielded as part of a national telephone survey (Krueger and Stone 2008).

In the next section of the paper, I will review several methods for capturing time use, with an emphasis on those that allow researchers to link activities to hedonic well being. I will describe how these methods work, and their relative advantages and disadvantages, both in terms of their validity—including their ability to reduce common response biases--and in terms of their feasibility for use in large-scale data collections, such as the HRS. Finally, I will discuss how the addition of time use measures could strengthen the HRS, by a) discussing new areas of research that could be addressed, and b) describing how some existing findings could be clarified through a better understanding of time use and well being.
Methods for capturing time use and well being Part I: Recall-based survey questions

The simplest way to measure time use is to simply ask respondents to estimate it directly—as in a question like “How many hours in a typical week do you spend on a commute to and from work?” These sorts of estimates are relatively easy to obtain, and can be correlated with other measures to answer various questions about the effects of longer versus shorter commutes (e.g., how does length of commute relate to a measure of global subjective stress?). However, most people do not have a ready response to this type of question, and therefore must create an estimate on the spot, based in part on their memories, and in part on their beliefs about what is “typical” for them. But how accurate are such estimates? Obviously, the answer depends somewhat on the nature of the activity in question. In the case of estimates of hours spent at work, for example, estimates could be fairly accurate, because these are activities that, at least for some people, vary little from day to day, and are measured in discrete, readily defined time intervals. However, relatively few activities that we may wish to measure show these properties--most vary from day to day, and are not necessarily attached to any fixed schedule. If, for example, one wanted to assess how much of a typical workday was spent engaged in various activities (meetings, administrative work, e-mail) the level of complexity increases rapidly—and, as argued below, accuracy diminishes.

In cases of variable, non time-fixed activities, the relatively simple method of directly estimating time use is prone to problems in both overall accuracy, and--even more problematically --biased estimation. Because people do not encode an hour by hour record of how they spend their time, they must construct estimates based on cues that they can remember—and these recollections are subject to various biases, such as the tendency to better remember intense experiences, as well as experiences that have occurred in the more recent past (e.g., Redelmeier, Katz et al. 2003) for reviews, see (Larsen and Fredrickson 1999; Smith, Brown et al. 2008)). Suppose a respondent is asked to estimate how many hours in a typical week they spend socializing with friends. The respondent must search his or her memory for episodes of this activity, estimate how long each episode lasted, sum the result over a week, and then decide whether to make some adjustment depending on a perception of whether this estimated week is “typical”. Several opportunities for inaccuracy present themselves here. First, in searching their memories, people are likely to recall vivid or intense (and therefore easily remembered), or recent occurrences of the activity in question. But recent experiences may not be typical, and vivid experiences, by definition, are not. In general, people are prone to use the availability heuristic—they tend to overestimate the frequency of easily remembered or imagined events (Tversky
and Kahneman 1973). This would lead to estimates of time use that are biased high—with respondents overestimating the amount of time spent in the activity in question.

Of course, not all time use questions of this sort ask about a “typical” week or other time period. But even a question that asks directly about activities over some period of time still requires remembering occurrences of the activity, estimating the length of these occurrences, and summing these lengths over the time period in question—and all of these judgments are subject to inaccuracy and potential bias.

Recall approaches for capturing well being

In many instances researchers are not merely interested in describing how much time people spend in various activities—they would also like to know something about the subjective experiences felt during these activities. Such information is crucial to understanding the hedonic value of different activities, which in turn allows for understanding of how different patterns of time use result in different experiences of emotional well being.

Traditionally, items assessing subjective experience ask people to recall their emotional experiences, impressions, or physical sensations in some situation (e.g., amount of pain experienced while doing housework, or walking), or over some specified period of time, such as the past week. For example, an item on the widely used SF-36 questionnaire asks: “During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?” Such items do not formally assess time use, in that they do not produce a quantified estimate of time spent in various activities, but they can be used to probe questions about how different activities affect subjective experience.

This method of using recall-based subjective assessment has many strengths, including being relatively inexpensive and easy to administer. Recall-based subjective assessments have generated a wealth of data about the subjective perceptions of well-being, and have been used in a variety of research contexts. In health research, items from recall-based subjective assessments are often the means for quantifying whether a treatment has measurable benefits for quality of life. Advocates of recall-based measures have even called for increased reliance on these types of measures for making policy decisions (e.g., Dolan 2008).

Despite these strengths of recall-based measures, psychological research suggests that caution should be used when deciding whether to rely on data obtained with recall-based methods. Studies show, for example, that subjective recall-based ratings are biased by situational factors such as momentary mood and personality (Diener, Suh et al. 1999; Schwarz and Strack 1999). Retrospective
reports of happiness, for example, can be influenced by momentary events or mild emotional boosts such as finding a dime right before filling out a survey (Schwarz and Strack 1999). In recalling subjective experiences, participants also tend to draw on information that is most readily available (Schwarz 1998; Schwarz and Strack 1999). Even seemingly minor factors, such as the introductory content or a preceding question on a survey, can greatly inflate the relationship between a sub domain of quality of life (e.g., health-related quality of life) and a more global measure (Smith, Schwarz et al. 2006). Individuals who were asked to rate marital satisfaction prior to overall life satisfaction, for example, showed an artificially strong correlation between the two, presumably because the easy availability of thoughts about marriage bias people’s judgments of their overall life satisfaction (Schwarz and Strack 1999).

Recall-based measures may also reflect patients’ expectations of well being as opposed to their experience of well being. People tend to supplement incomplete memories of emotional experiences using implicit theories to construct their responses, a phenomenon which has been described as a “theory driven recall bias” (Ross 1989). For example, an individual may inflate their levels of stress experienced during commutes to work because of a belief that commutes should be stressful—whether or not they had experienced high levels of stress (Ross 1989; Mahomed, Liang et al. 2002; Smith, Sherriff et al. 2006).

An additional limitation of these measures lies in their static, summary nature. Because they summarize experiences over some period of time, they lack information about fluctuations, patterns of change, etc, and do not provide the opportunity to look for immediate antecedents and consequences. In the next section of the paper, I will describe alternative techniques designed to assess time use and well being. These techniques are less reliant on memory—and thus less prone to the biases described above—and some of them provide for the possibility of looking at well being dynamically, capturing the natural within day fluctuations that accompany different types of activities.

Part II: Alternative approaches

End of day diaries

End of day written diaries have been profitably used in a variety of research areas for decades. They hold several advantages; they are self administered in real life settings, and assess immediate or very recent time periods, and therefore should be somewhat resistant to some of the biases that afflict traditional recall based survey items (Shiffman, Stone et al. 2008). Because they typically ask about experiences over the entire day, this method can readily be applied to research questions about time use. However, written diaries are still reliant on respondents
ability to recall and summarize their experiences over the day—and some research suggests that memory bias can affect time periods as short as a day or less (e.g., Kahneman and Redelmeier 1993). Diaries are also typically administered at night, which itself may impose some bias. Finally, because they also typically ask respondents to summarize or count their experiences or behaviors, they do not allow for the capture of dynamic changes in subjective experiences throughout the day.

Ecological Momentary Assessment

Ecological Momentary Assessment (EMA) refers to a class of techniques in which respondents provide subjective assessments of physical symptoms and psychological well being in real time, as they go about their daily lives (Stone and Shiffman 1994; Shiffman and Stone 1998; Shiffman, Stone et al. 2008). As such, they have several advantages over the other methods of assessment described above. First, because the assessments occur in—or close to—real time, they reduce the problematic effects of biased recollections. In addition, because they are assessed in real life situations, rather than in a clinic, or a research setting, they reduce potential unwanted context effects created by those settings. Finally, because assessments are repeated, they allow for consideration of the dynamics of time use and well being—including analysis of the immediate antecedents and consequences of fluctuations in subjective experiences. Thus, EMA techniques can be used to reliably capture time use—and the subjective emotional states that accompany different activities—but they do not rely as heavily on respondents’ ability to recollect their experiences.

In addition, because of their focus on immediate experience, EMA methods are ideally suited to assess momentary affect, which represents an important element of well being that is distinct from concepts like life satisfaction. Traditional survey items often assess concepts like life satisfaction, which represent not just hedonic well being, but also a cognitive appraisal component. EMA methods, on the other hand, are designed to capture momentary affective experiences, and thus can be used to assign specific hedonic values to different types of activities.

There are several approaches to EMA, but the term originally referred to the method in which the respondent wears or carries an electronic device for a period of time, such as a week, that prompts them at various times throughout the day to respond to a brief survey. The survey typically first asks respondents to indicate what they are doing, and can then assess their current level of any aspect of subjective experiences (e.g., mood, pain, stress, etc). This particular form of EMA is sometimes called experience sampling, because thin slices of different experiences are repeatedly sampled, providing a dynamic picture that, averaged
over time and across participants, can produce an accurate representation of experienced well being in various real life settings. Thus, from this type of data, researchers can compute average levels of variables of interest, and also explore peak and diurnal experiences. There are many possible schedules of prompts for responses. For example the device could be programmed to prompt the respondent randomly, many times per day, or on a fixed time interval (e.g., upon waking, then every 4 hours until bed time). Alternatively, the sampling frame could be event-based—the respondent could be instructed to input responses when they experience some symptom, like a pain flare up, or after they engage in an activity, such as exercise, or after taking a medication, etc. These strategies have different advantages and disadvantages; the random method, for example, avoids bias, but may miss important experiences. In addition, the choice of schedule has implications for how this method can be used to assess time use, which I will discuss at the end of this section.

A substantial advantage of these methods is that they do not rely on memories constructed after the fact—rather, they can capture affect directly, at the time and place experienced. Studies using EMA methods have revealed findings that either could not have been discovered using traditional survey measures, or contradict those measures. For example, in a study of fibromyalgia patients, recall-based pain measures showed improvement in the absence of any clinical intervention, whereas the real time assessments using an EMA method did not show the artifactual improvement (Williams, Gendreau et al. 2004). These methodological differences may be due to the fact that momentary assessments differ from memories of emotions and physical symptoms in systematic ways (Robinson and Clore 2002). For example, patients undergoing colonoscopy gave retrospective evaluations of their total pain during the procedure that were highly influenced by peak pain level during procedure, and by pain experienced during the last three minutes, as assessed by online pain measures (Redelmeier, Katz et al. 2003). Kahneman and colleagues have reported similar findings for the biasing influence of recent events on recalled experience of pain (Kahneman, Fredrickson et al. 1993).

Thus, measures based on EMA methods appear to benefit in terms of accuracy—but they also benefit in terms of the richness of data provided, in that they can better describe experiences and behaviors as they vary both across and within days. Numerous examples can be found in the literature (for a review, see Shiffman, Stone et al. 2008), and the evidence to date underscores the possibility that EMA may be especially valuable for understanding self-management processes, in terms of health-maintaining behaviors and in chronic illness. These newer measurement tools have the potential to provide a more complete and nuanced view of how behaviors change over time, and how factors such as mood and competing demands might negatively influence self-management.

DOI: 10.2202/1558-9544.1266
EMA methods and time use

It would be difficult to create a formal measure of time use from EMA methods in which slices of experience are intermittently sampled—we do not know what happened between sampling intervals. However, if the sampling frame is random, and either a sufficiently high number of experiences are sampled, or the sampling is done across a very high number of respondents, an accurate overall sense of time use could still be obtained. In addition, a slight modification in the protocol allows for a more comprehensive assessment of a day’s activities and experiences—respondents can be prompted to summarize their activities and affective experience since the last prompt. This hybrid approach has some of the advantages of experience sampling methods, and can be used to create a contiguous, comprehensive picture of how a respondent spent his or her day, along with affective ratings. But it also carries some disadvantages; most notably, a reliance on memory—however short the time frame—along with the requirement that experiences be summarized over that frame. In addition, the sampling intervals are random or arbitrary—rather than event based. Thus, if a prompt occurs after several different activities, the experiences must be averaged across them, and important dynamic information is lost. Likewise, if a prompt occurs half way through an activity—say, an emotional conversation—the rating of both the prior and subsequent interval contain only a part of this activity, potentially artificially reducing its apparent importance. It might be possible to design a protocol that would allow a participant to break down an interval into several activities and provide ratings of each, but such an approach will rapidly multiply the length—and therefore burdensomeness—of the method.

Summary

EMA methods for assessing time use and well being hold a number of advantages over traditional recall-based survey items, both in terms of accuracy and in the richness of data provided. However, EMA methods can be expensive, cumbersome, and represent a substantial burden to respondents. For these reasons, it is generally not feasible to employ EMA methods in large-scale national surveys, such as the HRS. In addition, some EMA methods, such as experience sampling may miss rare but important events, because the sampling intervals are random or arbitrary.

A potential alternative to EMA: Structured Diary Methods

Structured diary methodologies are designed to avoid problems of many recall-based measures, but can be considerably less burdensome than EMA methods.
(Kahneman, Krueger et al. 2004). A prominent example of this method is used in the American Time Use Survey, conducted by the Bureau of Labor Statistics, wherein participants follow a structured format in which they first divide the previous day into specific ‘episodes’ or events. They estimate the length of each event, describe them in terms of the type of activity (e.g., commuting to work, having a meal, exercising), and the setting in which this activity occurred. This produces a detailed snapshot of how the respondent spent the previous day.

Recently, this approach has been expanded to incorporate subjective affect ratings for each listed activity—the Day Reconstruction Method (DRM; (Kahneman, Krueger et al. 2004)). Thus, the DRM provides data that is similar to EMA data, in terms of the ability to dynamically track affective states as they relate to how time is spent, and the ability to capture the sequential nature of experiences. But it differs in at least two important ways. First, because the DRM focuses on events, rather than random or arbitrary intervals, it should be more likely to capture important experiences than experience sampling methods—even when they are relatively rare or of short duration. Second, because participants are rating the previous day, the DRM does rely, somewhat, on recall. However, the DRM exploits the fact that memory with respect to discrete events is much better than summaries memories of general affective states (Robinson and Clore 2002).

The structured format, wherein respondents first reconstruct the events of the previous day and then their subjective experiences of those events, increases the association between actual experience and what the respondent remembers, which reduces the chance that subjective biases can influence memory. In this way, the DRM can be relatively immune to some of the biasing factors noted above, such as the tendency to recall information that is congruent with a current mood state.

Psychometric studies support the validity of the DRM, and show that it can replicate findings from EMA studies that reveal counter-intuitive relationships between time of day and affect—relationships that have not been shown using recall-based methods (Kahneman, Krueger et al. 2004). Although the DRM method has been primarily used to measure emotional well-being, it was designed to be adaptable to a variety of research purposes, and is now being used to measure health related quality of life outcomes, thoughts, and behaviors (Krueger and Stone 2008; White and Dolan 2009).

A limitation of the DRM approach is the focus on a single day. Just as with EMA measures, the DRM can miss important but relatively rare occurrences that occur outside of its limited sampling range. Of course, structured diaries can be repeated over several, or many days, but such an approach may not be feasible in a large scale national survey like the HRS.
The DRM, time use, and well being

The structured format of the DRM breaks down a day’s events, along with an estimate of time spent in each activity. Thus, the DRM is the first diary method specifically designed to comprehensively assess time use in a way that is explicitly linked to subjective experience. In addition, the method holds many of the advantages of EMA approaches (reduced susceptibility to biased recall, the ability to examine moment to moment changes dynamically), but at a substantially decreased burden to participants and researchers, allowing it to be employed in large scale surveys (Kahneman, Krueger et al. 2004).

Completion of the full DRM with the original affective measures (12 items) requires about 45 minutes, on average. In two studies of chronic pain in older adults, we expanded the measure to also include measures of physical symptoms and impressions of social interaction and found that this did not substantially increase time to complete the measure (Smith and Murphy, unpublished data). As mentioned, the American Time Use Survey (ATUS) started in 1993, uses a similar structured diary format, collected via telephone interview, in a large scale national survey. In fact, the structured diary used in the ATUS survey shares many elements with the DRM (excluding affective well being), allowing for the potential for comparisons of time use patterns with a nationally representative sample. The DRM itself has also been fielded in large survey studies, including in a recent investigation of chronic pain in a community sample (Krueger and Stone 2008).

Forty five minutes represents a burden on respondents that is likely too great for many large scale surveys including the HRS. To decrease respondent burden, Krueger and Stone modified the DRM by gathering pain ratings on a subsample of the activities respondents listed, rather than gathering a comprehensive assessment of the previous day. By combining these sampled activities over a large number of respondents (approximately 4000), the authors were able to estimate the prevalence of experienced pain in the community, and the proportion of time spent in pain for various demographic groups, and during various activities.

Part III: How would time use/well being measures strengthen the HRS?

New research questions

The HRS utilizes a comprehensive set of measures regarding economic circumstances, health, family structure, and many other topics on a cohort
sampled repeatedly over the past 18 years. In general, we can think about how adding measures of time use and well being would illuminate a) how these variables affect how people use time vis-a-vis well being, and b) the potential meditational effects of time use and well being on subsequent outcomes like mental and physical health. The possibilities are too numerous to list here, but I will briefly list several general avenues of research that would be opened by the inclusion of such measures.

1) How does a decline in health status affect how people spend their time? How does it affect the hedonic value of activities? How do these changes themselves link to later outcomes such as overall well being, depression, economic status, subsequent health outcomes, and health behaviors?

2) Are different patterns of time use more or less associated with emotional resilience or adaptation in the face of declining health or other adversity?

3) How do life transitions—changes in employment status, in wealth, in marital status, bereavement, etc.—relate to changes in time use and experienced affect?

4) How does wealth relate to time use? Does this relationship change after the onset of a disability, or becoming a caregiver to an ill spouse? Can wealth ‘free up’ time for people in difficult circumstances in ways that buffer against the negative emotional impact of those circumstances?

Clarification of previous findings

In addition to opening various new avenues for research, employing methods that reveal the dynamics of well being and time use can help clarify existing research findings. Consider a recent study of caregivers of people with cognitive impairments. Numerous studies using traditional assessments of well being have demonstrated that being a caregiver of a cognitively impaired spouse or other relative can be stressful, and possibly damaging to one’s physical health. Poulin and colleagues utilized EMA to capture the experiences of caregivers as they went about their day (Poulin, Brown et al. in press). Electronic devices prompted participants several times a day, and asked them how much time in the past several hours were spent providing active care to their cognitively impaired spouse, and how much time was spent engaged in more passive, “supervisory” care. Time lagged analyses produced an intriguing pattern: positive mood actually increased after intervals in which more active care was provided. In contrast, negative mood was elevated after intervals containing more supervisory
care. These findings do not mean that being the caregiver of a cognitively impaired loved one is not difficult and stressful, but they do suggest that the act of providing active care may not be the aspect of this role that is harmful.

In a separate study utilizing HRS data, and using a more traditional estimate of caregiving, Brown and colleagues found that mortality risk over a seven year period declined as reported hours of care increased—in contrast to other studies that have suggested an increased risk of mortality as a function of caregiver status (Brown, Smith et al. in press). But from the EMA study of caregiving above, we can reasonably hypothesize that the estimates of reported hours of care in the latter study were biased high, and that the type of caregiving—the ratio of active to supervisory help—might determine whether time spent caregiving is helpful or harmful with respect to the health of the caregiver. Thus, a time use measure that accurately captured the relative amounts of these two distinct type of care could be utilized to refine our understanding of how caregiving affects the health of the caregiver, leading to a more precise understanding of the costs and benefits of providing care, and ultimately to more effective intervention designs.

Many other examples of findings that can be clarified by better time use measures could be described, but for the sake of brevity I will briefly mention just a couple more. A recent study using HRS data showed that the number of hours worked is associated with a decreased likelihood of exercising (Loh 2009). A more dynamic measure could reveal much more about the sorts of time use patterns that represent the biggest barriers to exercise. Finally, many HRS studies examine transfers of wealth from elderly parents to adult children. Measures of time use would allow for further exploration of the causes and effects of these transfers—time spent caring for a disabled parent, time engaged in positive or supportive interactions, as examples.

Conclusion

The HRS has tracked a nationally representative sample of older adults for nearly 20 years. The comprehensive assessments of economic, demographic, and psychosocial variables provide a comprehensive portrait of this cohort as they transition through retirement and into old age. A better understanding of the emotional experiences of this cohort as they engage in various life activities, would allow for leveraging the enormous amount of information already collected as part of the HRS, to open up numerous new areas of inquiry.

Relatively intensive methods such as EMA, with electronic diaries, are not well suited for large scale survey efforts such as the HRS. The more simplistic method of asking participants directly suffers from numerous well established biases, and does not provide the ability to track activities and emotional well
being dynamically—that is, in a way that allows for the examination of immediate antecedents and consequences of different types of activities and emotional states. The Day Reconstruction Method, although based on recall, provides many of the advantages of EMA methods in a format that is much less burdensome to respondents, and has already been used in several large-scale surveys. While to date, the DRM has primarily been used to capture mood, the format allows for the measurement of many aspects of subjective experience, including physical and psychological symptoms, cognitions, and behaviors.

Challenges to implementing time use measures include burden to respondents, and complexity of administration; however, there are methods for decreasing burden imposed by both EMA and structured diary formats, if one is willing to accept either fewer samples across time, fewer constructs to measure, or both. As always, researchers must weigh the costs of making such compromises against the benefits of being able to administer the methods in large scale survey studies. A related issue is whether such methods provide valid data in much older populations, especially where there is some cognitive impairment. To date, I am not aware of studies that directly address this issue. However, as I have reviewed, the standard survey format imposes its own sort of cognitive burden—especially with respect to memory. It is worth noting that to the extent that these methods reduce reliance on memory, they may turn out to provide more valid data among those with mild cognitive impairment than the more traditional recall-based formats.

References


DOI: 10.2202/1558-9544.1266


