

Developments in the Day Reconstruction Method and Related Methods: Review and New Directions

Los Angeles, California
January 8-9, 2015

MEETING SUMMARY

A conference supported with funding by the National Institute on Aging to the Edward R. Roybal Center for Translation Research in the Behavioral and Social Sciences of Aging at the Center for Research on Experience and Wellbeing, Princeton University

In collaboration with the Organisation for Economic Co-operation and Development
and
Hosted by the University of Southern California Dornsife Center for Self-Report Science

Arthur A. Stone, University of Southern California, Program Chair
Conal Smith, Organisation for Economic Co-operation and Development, Program Co-Chair

Revised May 4, 2015

FOR ADMINISTRATIVE USE ONLY



This meeting summary was prepared by Chandra Keller-Allen, Rose Li and Associates, Inc. under a subcontract to Princeton University (Purchase Order PRINU-0000023396) with federal funds from the National Institute on Aging, National Institutes of Health. The statements, conclusions, and recommendations contained in this document reflect opinions of the meeting participants and are not intended to represent the official position of Princeton University, the Organisation for Economic Co-operation and Development, the University of Southern California, the National Institute on Aging, the National Institutes of Health, or the U.S. Department of Health and Human Services. Review of earlier versions of this meeting summary by the following individuals is gratefully acknowledged: Somnath Chatterji, Michael Daly, Angus Deaton, Sarah Fleche, Patricia Houle, Thomas Kubiak, Rose Maria Li, William Michelson, Talya Miron-Shatz, Lisbeth Nielsen, Jacqui Smith, Arthur A. Stone, Samuel Thomas, Ruut Veenhoven, and Yoshi Yamamoto.

TABLE OF CONTENTS

ACRONYM DEFINITIONS	iv
EXECUTIVE SUMMARY	v
Overview	v
Future Research Considerations	v
Improve Ability to Inform Policy and Advise National Statistics Organizations.....	vi
Create Better Documentation.....	vi
Expand Research in New or Understudied Areas.....	vi
Systematically Assess Trade-Offs	vii
Summary of DRM Concepts	vii
CONFERENCE SUMMARY	1
Introduction	1
Perspectives from the Development Team	2
Substantive Developments.....	3
DRM in the American Time Use Survey (ATUS).....	3
Experienced Wellbeing in France and the United States	4
Time Use in Canada: A 24-hour Recall Collection	5
Method Effect of DRM: Users Become Happier.....	6
Fatigue and Mood Assessed with the DRM and EMA.....	7
Affect in Time Use Surveys	8
DRM and Self-Control.....	8
Discussion of Substantive Developments	9
Method versus Content.....	9
Issues of Scale.....	10
U-index for the Unemployed.....	10
Using Weights.....	10
Biomarkers	10
Individual versus Population Levels	11
Methodological Developments.....	11
Collecting DRM Data	11
Single Day DRM—Suitability for Survey Work	12
The Partitioning of the Day and the Heterogeneity of Daily Experience	14
DRM Aggregation Issues.....	16
Short Versions of the DRM: Ready for Prime Time?	16
Aggregation over the Day versus Mixed Methods Estimating and Weighting of the DRM	19
Do We Know If Reinstantiation Is Important?.....	19
EMA versus ATUS Wellbeing	21
Measuring the Effect of Thoughts on Daily Wellbeing.....	23
Day Reconstruction and Linkage to Other Data.....	24
Integrated Assessment Framework DEDIPAC	25
Discussion of Methodological Developments.....	25
Communicating Survey Benefits to Respondents.....	25
Multiple or Secondary Activities	25
Monitoring Effects.....	26

Trade-offs 26

The Role of the DRM in Informing Policy.....27

Future Research Agenda Considerations.....28

 Improve Ability to Inform Policy and Advise National Statistics Organizations..... 29

 Create Better Documentation..... 29

 Expand Research in New or Understudied Areas..... 29

 Systematically Assess Trade-Offs 29

 Summary of DRM Concepts 30

 List of Concepts 30

CONFERENCE AGENDA..... 32

PARTICIPANT LIST 34

ACRONYM DEFINITIONS

Acronym	Definition
ALP	American Life Panel
ATUS	American Time Use Survey
BLS	Bureau of Labor Statistics
BSR	Behavioral and Social Research
CAPI	computer-assisted personal interviewing
CATI	computer-assisted telephone interview
CREW	Center for Research on Experience and Wellbeing
DEDIPAC	Determinants of Diet and Physical Activity
DRM	Day Reconstruction Method
DUST	Disability and Use of Time
EDT	Enquête Emploi du temps
ELSA	English Longitudinal Study of Ageing
EMA	Ecological Momentary Assessment
ERM	Event Reconstruction Method
GSS	General Social Survey
HRS	Health and Retirement Study
HWB	Gallup Healthways Wellbeing Index
INDEPTH	International Network for the continuous Demographic Evaluation of Populations and Their Health
NIA	National Institute on Aging
OECD	Organisation for Economic Co-operation and Development
P-ATUS	ATUS conducted by Gallup for Princeton University
PATS	Princeton Affect and Time Survey
PSID	Panel Study of Income Dynamics
ROBUST	Research on Well-Being and Use of Time
SAGE	Study on global AGEing and adult health
SWB	subjective wellbeing
USC	University of Southern California
WHO	World Health Organization

EXECUTIVE SUMMARY

Overview

Since the development of the Day Reconstruction Method (DRM) a decade ago, investigators from around the world have used the method or variants thereof to understand how an individual experiences a single day. This meeting brought together prominent experts from multiple fields, including those who were involved in the creation of the DRM, those who have developed and implemented related measures, and those who have used the DRM or adapted approaches to measure experienced wellbeing in their own research. The goal of the conference was to inform the development of a future research agenda to better understand and improve the DRM and to provide guidance for national statistics organizations on substantive and methodological issues for measuring experienced wellbeing, particularly when evaluating trade-offs between using the full DRM and adapted versions that aim to reduce respondent burden.

Following an overview from members of the DRM development team, invited speakers gave presentations on both substantive and methodological developments related to the DRM and DRM hybrid approaches. Presentation topics included DRM hybrid versions used in national studies in the United States, France, and Canada, and in other large-scale studies such as the Health and Retirement Study (HRS), the Study on global AGEing and adult health (SAGE), and the English Longitudinal Study of Ageing (ELSA); linking DRM data to other types of data such as biomarkers; variations of data collection methods across multiple hybrid versions; methodological issues such as aggregation and weighting; the importance of reinstantiation; and how data from the DRM and its variants can inform public policy.

Discussions throughout the meeting aimed to identify key unresolved issues and items ripe for future research. Meeting chair Arthur Stone challenged the group to think critically about what would define success for the DRM (and related methods), assuming momentary experience is the gold standard for measuring experienced wellbeing. Meeting co-chair Conal Smith encouraged participants to consider what information is necessary to provide informed guidance specific to the needs and interests of national statistics organizations.

Future Research Considerations

The DRM was designed to be an efficient measure of wellbeing closely related to experience for use with large samples. Since its development, the full DRM has been used in numerous research studies and adapted into several hybrid versions for different purposes, scales, and populations. The DRM is valuable for a host of interdisciplinary scientific purposes that aim to measure and alleviate misery or increase happiness and is particularly well suited for evaluating programs or small-scale policies, such as clinical interventions or end-of-life decision making. In order to best serve these purposes, future research is needed to further develop the DRM and understand the implications of and uses for its abbreviated variants.

Several methodological issues were raised throughout the meeting concerning the trade-offs between ecological momentary assessment (EMA) methods, the full DRM, and shorter adapted DRM versions and the implications of these trade-offs for national statistical organizations. EMA cannot be a gold standard, if only because a lot of activities are inaccessible by the method; many individuals cannot answer during their work hours (e.g., teachers, bus drivers, or brain surgeons). EMA can also prove too burdensome for large samples. In these ways, the DRM can be better than EMA. Conference participants offered suggestions for future research on the DRM both to understand and improve the method for a variety of purposes and to inform national statistical organizations of the substantive and methodological issues that merit consideration.

Improve Ability to Inform Policy and Advise National Statistics Organizations

- Identify the key features needed in a DRM hybrid method appropriate for national statistics organizations.
- Determine the purposes and circumstances under which it is appropriate for a national statistics organization to conduct a full DRM on a representative subsample versus a shorter hybrid DRM on the representative full sample.
- Conduct research to determine how many and which affect dimensions to measure in national studies.
- Evaluate the best way to aggregate time use and affect data (i.e., the U-index) to inform policy.

Create Better Documentation

- Create a simple yet detailed guide or handbook that describes how DRM data can and should be used, with particular attention to weighting methods.
- Create a complete methodological manual that includes comparisons of DRM and DRM hybrid methods.

Expand Research in New or Understudied Areas

- Consider a broad range of biomarkers and their relationships to the DRM.
- Analyze time use and affect data of understudied subgroups.
- Create better measures of or opportunities to collect data on multitasking or secondary activities.
- Use HRS data to examine the extent to which experienced wellbeing predicts longevity.
- Conduct cognitive interviews, experimental studies, focus groups, and secondary data analyses to examine the differences in how respondents delineate episodes in their day and how these vary by subgroup, instrument instructions, or other factors and the implications of these differences.
- Determine whether alternative metrics, such as the U-index, can be used to derive new metrics from DRM-type data to provide insights into various outcomes of interest (e.g., pain).

Systematically Assess Trade-Offs

- Systematically assess how DRM hybrid approaches, such as the American Time Use Survey (ATUS) and the HRS wellbeing module, work for a range of measures and whether the standard errors are acceptable. This could possibly be done with a simulation study.
- Conduct systematic comparisons of the EMA and DRM measures and compare reliability, validity, accuracy, etc.
- Empirically test and compare multiple DRM versions using various levels of reinstatement and EMA.
- Test and define the essential components of the DRM.
- Assess how each trade-off in the DRM-like methods compares to the full DRM and identify the implications for the types of research questions that can be addressed with each method.
- Test the full DRM and one or more DRM hybrid methods (and possibly EMA) in the same population to study the relationships and trade-offs. There might be opportunities to add onto existing studies to make these comparisons.
- Define a set of criteria by which DRM hybrid methods should be measured to determine how accurate, effective, and useful the method is for what purposes and populations.
- Determine whether a single rating of affect per episode is adequate to accurately measure experienced affect and document what is known about how a respondent chooses a rating (e.g., dominant mood, an average).

Summary of DRM Concepts

Numerous concepts—such as mode, length, selection of episodes, affect ratings, extent of reinstatement, and periodicity—should be tested to understand the following questions:

- What implications does each of these choices have for the results?
- Which of these variables is required to consider it a DRM method?
- Which variables are suitable for what purposes?
- What package or possible combinations of packages are appropriate to recommend to national statistics organizations?

CONFERENCE SUMMARY

Introduction

Since the development of the Day Reconstruction Method (DRM) in 2004 by Princeton University's Center for Research on Experience and Wellbeing (CREW), dozens of investigators from around the world have used the method or its variants as a means to capture data about how an individual experiences subjective wellbeing associated with episodes of activity actually experienced in the course of a recent or current day. At the time of this meeting, a Google Scholar search on the phrase, "Day Reconstruction Method," yielded 1,920 results, attesting to the widespread adoption of the method. Despite the impressive uptake of the methodology, there has been no systematic effort to catalog the achievements, concerns, and limitations of the method. A review of the strengths and limitations of the DRM is timely, given that large-scale time use surveys are increasingly including measures of experienced wellbeing. These include DRM-derived methods such as the experimental wellbeing module in the American Time Use Survey (ATUS)¹ and the French Enquête Emploi du temps 2009-2010.²

This meeting brought together experts from many fields of study—economics, psychiatry, decision science, social psychology, sociology, cognitive psychology, and health psychology—who have used the DRM or related methodologies, to discuss the strengths and limitations of the DRM and opportunities for future research. The meeting was organized by substantive and methodological topics of concern to researchers using the DRM. The goal was to constructively inform the development of a future research agenda.

Michael Quick, Interim Provost and Senior Vice President of Academic Affairs at the University of Southern California (USC), welcomed the meeting participants and noted that the DRM is one of the most influential methods ever developed for assessing experienced wellbeing. In addition to celebrating the widespread use of the DRM over the past decade, this conference provides a platform to assess the implications of the DRM for improving society in health and wellness. It is hoped that the meeting participants will broadly document the strengths and weaknesses of the DRM and offer ideas for future applications. Interdisciplinary methods such as the DRM hold tremendous potential for informing public policy to address complex societal challenges.

Meeting chairs Arthur Stone and Conal Smith added their appreciation for the time and expertise of the conference participants and encouraged vigorous discussion. Conal Smith observed that the Organisation for Economic Co-operation and Development (OECD) has focused on measuring subjective wellbeing (SWB) for the past 5 years. In 2010 when the OECD began a program to encourage national statistics organizations to measure SWB, three countries regularly published life satisfaction data—Canada, Israel, and New Zealand. Currently, 32 of the 34 OECD countries regularly collect measures of life satisfaction, and two countries—

¹ See <http://www.bls.gov/tus/wbdatafiles.htm>.

² See <http://www.insee.fr/fr/methodes/default.asp?page=sources/ope-eng-emploi-du-temps-edt-2009.htm>.

the United States and France—measure experienced wellbeing. A third country, Canada, is pilot testing methods for measuring experienced wellbeing. National statistics organizations must justify measuring SWB and need clear guidance on the related methodological issues for measuring experienced wellbeing, particularly when evaluating the trade-offs between the benefits of using the full DRM versus adapted versions that reduce participant burden.

The Division of Behavioral and Social Research (BSR) at the National Institute on Aging (NIA) supports behavioral and social research in economics, demography, epidemiology, psychology, and other social science domains that focus on processes at the individual and societal levels that keep people healthy as they age, which requires appropriate methods for measuring SWB. Lisbeth Nielsen noted that the NIA has supported the development of the DRM and its adaptation for the ATUS through its Roybal Centers program, and the Institute is eager to be informed by a future research agenda for these types of methods.

Perspectives from the Development Team

Members of the development team, including Stone, Alan Krueger, Norbert Schwarz, and David Schkade offered their reflections about the early development of the DRM and their experiences collaborating with each other and with Daniel Kahneman (who was unable to attend the meeting). Ecological momentary assessment (EMA), considered the gold standard for measuring experienced wellbeing, was too burdensome for large samples; at the time it cost \$800 per subject for an EMA study. The team set out to develop an efficient measure of wellbeing closely related to experience for use with large samples.

During the early 2000s, the team began to design and test iterations of the DRM instrument, often using homogeneous convenience samples (e.g., teachers, nurses) to minimize variance. At Schwarz's urging, a key feature of the DRM was reinstantiation of the previous day into working memory by producing a diary of time use documenting a sequence of episodes. Reinstantiation was followed by a series of questions to characterize each episode, including how the respondent felt in multiple affect dimensions. Schkade remembered that during early data collection several respondents had emotional reactions to the reinstantiation process, which gave him confidence in their responses.

Many of the team's initial hypotheses about wellbeing were not supported by early findings, and the development process yielded new research questions, which the team then iteratively pursued. In 2004, the development team published its first article on using the DRM to measure daily life experience.³ Schkade noted that readers seemed most interested in the activities and corresponding affects documented, such as child rearing and commuting, whereas the development team was focused on the instrument itself.

³ Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., & Stone, A.A. (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science*, 306, 1776-1780. Available from <http://www.sciencemag.org/content/306/5702/1776.full?ijkey=VrATbz271RslA&keytype=ref&siteid=sci>.

The DRM demonstrated the distinctions between experienced and evaluative wellbeing. For example, the activity of parenting rates differently when asked in an evaluative versus experienced manner. Several of the hypotheses that the team thought it was testing (e.g., hedonic treadmills on evaluative judgment) were not confirmed by experienced wellbeing data. Over time, the DRM was shown to reasonably approximate more time intensive experience sampling methods and reduce respondent bias.⁴

DRM-related methods could be used to influence social policy. For example, the U-index, which indicates the fraction of time people spend with the strongest emotion being negative could be a guide for public policies focused on alleviating misery, but requires further exploration (i.e., resolving issues of scale).⁵

Krueger reported on efforts that followed the creation of the DRM to design a module appropriate for national surveys, including the Princeton Affect and Time Survey (PATS), questions that were added to a World Health Organization (WHO) survey, and the creation of the ATUS experienced wellbeing module.⁶ These developments were facilitated in part by Richard Suzman, Director of the Division of Behavioral and Social Research at the NIA.

Stone challenged participants to think critically about what it would mean for the DRM and its related methods to be considered to be achieving its aims. The DRM is thought of as a wellbeing method; however, it can also be a generic approach to obtaining information efficiently in different domains. DRM is a methodology linked to wellbeing research that could be applied to multiple substantive areas such as pain, symptomology, social interactions, and unemployment.

Substantive Developments

DRM in the American Time Use Survey (ATUS)

Alan Krueger, Princeton University

Suzman was instrumental in working with the Bureau of Labor Statistics (BLS) to implement the ATUS wellbeing module, and data are now available for 2010, 2012, and 2013.⁷ In the module, after documenting time use for the full previous day, respondents are given three randomly selected activities from their time diary and are asked to rate their affect during each of the three selected activities on a scale of 0-6 for happy, tired, stressed, sad, pain, and meaningful.⁸ The module also includes questions about health status and pain medication.

⁴ Kahneman, D., & Krueger, A.B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives*, 20, 3-24.

⁵ See Kahneman and Krueger (2006).

⁶ See <http://krueger.princeton.edu/pages/atus-subjective-well-being-module>.

⁷ See <http://www.bls.gov/tus/#data>.

⁸ See the full questionnaire at <http://www.bls.gov/tus/wbmquestionnaire.pdf>.

The ATUS wellbeing module has yielded a large dataset available to researchers. The BLS weights required to analyze these data are complicated and challenging to use. Future activities include more aggressively promoting the availability and use of these data and providing clear and simple explanations and demonstrations of using the BLS weights in analysis.

Experienced Wellbeing in France and the United States

Sarah Fleche, Centre for Economic Performance, London School of Economics

As of 2014 only two countries—France and the United States—have included measures of experienced wellbeing in official time use surveys. Canada plans to do so in the future. None of the three countries is using the same approach to measurement. The different methods have implications for the data collected, and researchers need to better understand the strengths and weaknesses of the different methodological approaches.

As noted, the ATUS wellbeing module randomly selects three time intervals from the previous day's diary and asks for ratings on six dimensions of affect for each selected activity. The ATUS sample size is 12,829, and the sample is split evenly between weekdays and weekend days. In contrast, France's Enquête Emploi du temps (EDT) randomly selects a subsample of respondents to record a unidimensional affect (pleasant versus unpleasant) for every activity in the time diary for the previous day. The sample size is 2,638 diaries (1,355 individuals): two diaries per respondent, one about a weekday and one about a weekend day. See Table 1 for a summary of the key differences.

Table 1: Key differences between ATUS and EDT design

	Random Selection	Affect Rating	Sample Size	Summary
ATUS	Of subsample of episodes	Six dimensions	12,829 respondents, half reporting weekdays and half reporting weekend days	Multidimensional affect on a sample of episodes for all respondents
EDT	Of subsample of respondents	Unidimensional	1,355 respondents, each reporting both a weekday and a weekend day	Unidimensional affect on all episodes for a sample of respondents

The average number of episodes described for the Wellbeing Module per respondent is 2.98 in the ATUS and 18.89 in the EDT. The total number of wellbeing measurements is 27,485 in the ATUS and 29,480 in the EDT. One advantage of the EDT is the ability to evaluate the intra-episode variance of wellbeing (from 0 to 4.24 during one episode). The average observed wake time for participants is about 3 hours in the ATUS Wellbeing Module and about 18 hours in the EDT.

Fleche and Conal Smith are replicating a study that examined experienced wellbeing in Rennes, France, and Columbus, Ohio, using data from the ATUS and the EDT.⁹ They organized the activities into six broad categories (work/commute, compulsory activities, active leisure, passive leisure, eating, and other). The analyses presented refer to focal activities, and these data were re-weighted by day of the week to be representative of a random day.

The ATUS offers the advantage of being able to examine specific emotional states. However, the EDT offers six times as many episodes for the same sample size as does the ATUS. The EDT also allows for exploration of secondary activities and the change in wellbeing over time during one activity. Findings related to passive leisure were problematic in both the EDT and the ATUS.

Time Use in Canada: A 24-hour Recall Collection

Patricia Houle, Statistics Canada

The Canadian General Social Survey (GSS) Program collected time use data in 1986, 1992, 1998, 2005, and 2010. There are four versions of the instrument: English and French via either computer-assisted telephone interview (CATI) or Electronic Questionnaire. Individuals aged 15 and older are randomly selected from private households in the 10 provinces using a telephone sampling frame that includes cell phone numbers. A specific reference day is assigned to the respondent to report daily activities for a 24-hour period; the data are collected over a 12-month period so as to obtain information from all seasons.

For the CATI, interviewers code the reported activities at the time of data collection. There were 85 activity codes in 1986 and 264 codes in 2010, which was too many for the Electronic Questionnaire version. Therefore, many activities were combined for 2015, resulting in 72 codes. Also, the time interval will be changed from 5 to 10 minutes because past data indicated that activities rarely lasted fewer than 5 minutes. The survey asks about respondents' primary activity (duration, location, with whom, and then what) and simultaneous activities.

Historically, SWB has been measured in Canada through stylized life satisfaction questions. The 2005 time use survey included an enjoyment question at the end of the diary. After consultation with wellbeing experts, Canada has decided to use the French scale without the followup questions on affect to measure SWB in 2015. The same SWB question will be asked for two randomly selected time periods: "On a scale of -3 to +3 where -3 means unpleasant and +3 means very pleasant, at [random time] how would you rate the activity you were doing?"

The core survey will be administered from April 7, 2015, to April 6, 2016, using a mixed mode collection approach (CATI and Electronic Questionnaire). A mixed mode effect study will be completed as well. The data will be released in the summer of 2017.

⁹ Krueger, A.B., Kahneman, D., Fischler, C., Schkade, D., Schwarz, N., & Stone, A.A. (2009). Time use and subjective well-being in France and the U.S. *Social Indicators Research*, 93, 7-18.

Method Effect of DRM: Users Become Happier*Ruut Veenhoven, Erasmus University Rotterdam*

One reason the DRM was developed to measure experienced wellbeing was to overcome the measurement bias found with evaluative life satisfaction questions. However, the DRM has a meta effect on the user: repeated use of the DRM can raise happiness. This is an important feature to understand when using the DRM in effect studies. Veenhoven reported that the change in happiness for occasional use of the Happiness Diary using the DRM approach is the same as winning the lottery.

Veenhoven uses the DRM as part of an online happiness intervention called the Happiness Indicator.¹⁰ The conceptual framework of the intervention is that users who are more aware of their affect make better choices that lead to being happier, which in turn leads to better health and fewer doctor visits. The DRM in this intervention is used for raising awareness.

The website tools include the Happiness Comparer and the Happiness Diary, which both provide instant feedback to the user by comparing the user's level of happiness with the average of all participants, with people like the user, and over time. The comparer asks the user two questions—"How happy do you feel today?" and "How happy did you feel over the past month?"—which are answered on a scale of 0 to 10 accompanied by smiley faces. The first question aims to focus the user's attention on the affective side of happiness.

The Happiness Diary asks the user to list yesterday's activities, including start/stop times, location, with whom, and secondary activities. The user then rates mood on the same 0 to 10 scale for each activity. These results can also be compared to the results of other participants. The Happiness Diary raises the user's awareness and displays how the user could feel based on reports of others' happiness. The program generates a duration weighted average affect.

Veenhoven conducted an effectiveness study of 50,000 participants who used the tools since January 2011. Of these participants, 3,018 used the DRM diary at least once and 1,653 used the DRM diary twice or more. The user's rating on the 0 to 10 scale of happiness in the past month was the dependent variable. The effect of frequent DRM use was a 1.4 percent gain in happiness. Repeat use of the DRM resulted in diminishing additional effects. The effect was stronger for people who tend to be less happy on average. The results were similar across age, gender, income, and education and were unaffected by use of the Happiness Comparer tool.¹¹

The findings of this effect study might have implications for DRM use in research. The benefits to users can be explained to potential respondents as a tool for recruitment. Researchers need to be aware of possible positive bias. For example, if the DRM is used to assess the effect of psychotherapy, then a 1.5 percent increase in happiness could be due to using the DRM, not

¹⁰ See <http://www.gelukswijzer.nl/> or <http://www.happinessindicator.com/>.

¹¹ A more detailed description of the effect study presented can be found at http://www2.eur.nl/fsw/research/veenhoven/Pub2010s/HappinessIndicator-Effect_DRMconferenceLA2015.pdf.

the intervention. Correction of this bias is possible by assessing the effect of the DRM in the population and then adjusting the effect.

Schwarz suggested that the results are impacted by selection bias because users were volunteers. Veenhoven agreed that there is selection bias, and its effect has yet to be analyzed. The tools could be tested with the general public to assess this issue.

Fatigue and Mood Assessed with the DRM and EMA

Yoshi Yamamoto, University of Tokyo

Kahneman and colleagues (2004) posited that the DRM is a reliable substitute for EMA to measure subjective symptoms.¹² How well the DRM can trace changes in fatigue and mood over time is unknown. Yamamoto and colleagues conducted a systematic comparison of momentary (EMA) and retrospective (DRM) methods for fatigue and mood and demonstrated that the two methods are reasonably comparable.¹³ Behavior (as measured by activity, location, companion) and subjective symptoms (as measured by Visual Analogue Scale for fatigue and depressive and anxious mood scales) were assessed with EMA and the DRM in the same subjects. No significant differences were found between the mean or the variability estimated by EMA and the DRM. Correlations between EMA and the DRM for fatigue, depression, and anxiety were low, yet statistically significant.

In another study, Yamamoto and colleagues examined the covariation of depressive mood and locomotor dynamics in healthy adults using EMA.¹⁴ Locomotor activity was continuously measured with an actigraph. There were positive correlations among momentary symptoms. Increased intermittency of locomotor activity (i.e., reduced activity with occasional bursts) appeared concurrently with a worsening of depressive mood; this association was statistically significant regardless of group differences in age, lifestyle, and occupation. Yamamoto also showed a preliminary result of an additional analysis of data collected in the first study that such an association between mood and locomotor activity was not significant when using DRM.

The overall mean and variability of momentary fatigue and mood states as well as changes over time as measured by EMA can be assessed with the DRM. Yamamoto tentatively concluded that the DRM can capture changes in subjective symptoms over time with lower respondent burden than the EMA. Furthermore, combining locomotor activity data with the DRM provides additional information on physical symptoms and mood states that is likely captured by the EMA.

¹² Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., & Stone, A.A. (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science*, 306, 1776-1780.

¹³ Kim, J., Kikuchi, H., & Yamamoto, Y. (2013). Systematic comparison between ecological momentary assessment and Day Reconstruction Method for fatigue and mood states in healthy adults. *British Journal of Health Psychology*, 18, 155-167.

¹⁴ Kim, J., Nakamura, T., Kikuchi, H., Sasaki, T., & Yamamoto, Y. (2013). Co-variation of depressive mood and locomotor dynamics evaluated by ecological momentary assessment in healthy humans. *PLoS One*, 8, e74979.

Affect in Time Use Surveys

William Michelson, University of Toronto

The primary role of time use measurement over the past century has been to produce objective and comprehensive documentation of how people spend their time. However, a full understanding of time use also requires knowing what people think and feel about what they are doing. In the early years of time use measurement, prominent scholars imputed motives to time spent in specific activities that reflected their home disciplines, but these do not necessarily reflect the affects experienced by individuals.

Affective information must be assessed directly. Affect can be observed at three levels: (1) individual episodes of behavior (hedonic); (2) life satisfaction (macro); and (3) in between—not hedonic (e.g., time pressures or perceived stress). The 2009 Stiglitz Commission Report urges the use of both micro and macro affect measures in national time use surveys.¹⁵ However, the comparability of the measures of affect, especially those of different levels of scale, is unknown. Another unanswered question is whether there are heuristic non-hedonic variables that can capture the essence of affect found at both micro and extreme macro levels. Definitive answers to these questions are hindered by time use studies that lack similar, robust, and comprehensive affective content covering diverse levels of scale.

Michelson conducted exploratory work to examine data from two time use studies—the Employed Mothers Study (1979) and Canada’s General Social Survey (GSS; 2005)—that differ in scale, representativeness, and selection of affective measures. His analyses suggest that affective measures within micro or macro levels produce similar or complementary results. Associations between measures at opposite levels of scale are weak in comparison. Some measures of general feelings of time pressure appear to capture affective feelings generated by hedonic episodes and life satisfaction (e.g., phenomena having to do with time pressure). These *hinge* variables that bring together feelings from opposite levels of scale suggest underlying considerations that are bound together. Future research could include increasingly wider ranges and useful choices of affective measures when time use data are collected.

DRM and Self-Control

Michael Daly, Behavioural Science Centre, University of Stirling

Self-control involves the effortful regulation of thoughts, feelings, and behaviors to resist temptations and inhibit impulsive actions. A meta-analysis by de Ridder and colleagues found that self-control is positively correlated with favorable educational, work, social, and health outcomes.¹⁶ Childhood self-control predicts adult outcomes independent of socioeconomic

¹⁵ Stiglitz, J.E., Sen, A., & Fitoussi, J.P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. Retrieved January 28, 2015, from <http://www.stiglitz-sen-fitoussi.fr/en/index.htm>.

¹⁶ de Ridder, D.T.D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F.M., & Baumeister, R.F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review*, *16*, 76-99.

status and intelligence.¹⁷ There is a positive link between self-control and wellbeing across 16 different comparisons and with substantial heterogeneity. High levels of self-control are linked to higher life satisfaction and momentary wellbeing,¹⁸ less variability in emotion, fewer emotional extremes, and greater emotional stability. There has been little work using the DRM to examine intra-individual standard deviation of affect.¹⁹ Work that has been done shows that individuals who have variable positive emotions tend to also have variable negative emotions.

Daly's current study aims to examine the relationship between trait self-control, affect, and affect variability measured with the DRM; identify links between self-control and stress biomarkers; and test whether emotional patterns and health behaviors explain linkages between self-control and psychobiological functioning.²⁰ Data were collected from 198 students in an online DRM using the affect adjectives happy, calm, comfortable, affectionate, interested, confident, impatient, depressed, stressed, and irritated. The majority of respondents also provided heart rate and cortisol data.

Results indicated that self-control is associated with a higher positive affect and lower affect variability and therefore may have emotional and psychological benefits. Better self-control was linked to a lower resting heart rate and a steeper cortisol slope. Mediation analyses showed that affect variability explained the link between better self-control and a steeper cortisol slope. Smoking explained the link between self-control and resting pulse and partially explained the link to heart rate variability. Daly concluded from these results that psychological traits could be linked to affect and biological functioning using the DRM.

Future work will include examining the long-term link between childhood self-control and wellbeing in adulthood, linking time discounting and wellbeing using the DRM in the English Longitudinal Study of Ageing (ELSA), and integrating desire, goal conflict, resistance, and enactment into the DRM.

Discussion of Substantive Developments

Method versus Content

Nielsen reiterated Stone's introductory comments that the DRM is a methodology, but the content of domains or adjectives can vary depending on the research question(s). Researchers interested in social lives or symptomology or pain could have different sets of affect adjectives. It is important not to wed the DRM to a particular list of affect adjectives, although it might be important to outline the necessary characteristics of the set of adjectives needed for the DRM.

¹⁷ Moffitt, T.E., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS*, *108*, 2693-2698.

¹⁸ Hofmann, W., Luhmann, M., Risher, R.R., Vohs, K.D., & Baumeister, R.F. (2013). Yes, but are they happy? Effects of trait self-control on affective well-being and life satisfaction. *Journal of Personality*, *82*, 265-277.

¹⁹ Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*, 662-676.

²⁰ Daly, M., Baumeister, R.F., Delaney, L., & MacLachlan, M. (2014). Self-control and its relation to emotions and psychobiology: Evidence from a Day Reconstruction Method study. *Journal of Behavioral Medicine*, *37*, 81-93.

Issues of Scale

Schkade provided cautionary comments about the use of a unidimensional scale for rating affect. Canada and France use a bipolar scale with 0 in the middle and ratings of positive or negative numbers (i.e., -3 to +3). Kahneman and Schkade were unable to detect loss aversion with this type of scale. Respondents first decide whether their affect is positive or negative and then decide the intensity. However, the units of intensity are not the same for positive and negative affect. This is not a problem for calculating the U-index because the sign, not the intensity, is important. But for analyses involving means or aggregates, the different units of scale for positive and negative affects is a problem. Schkade encouraged researchers to consider different scales or ask for two separate reports of affect, one positive and one negative.²¹

U-index for the Unemployed

Schwarz, Krueger, and Fleche discussed interesting findings concerning the U-index for the unemployed. Schwarz remembered a German dataset that demonstrated that the unemployed replace unpleasant activities with somewhat pleasant ones, but generally enjoy the activities less. Krueger found that the U-index is high for the unemployed. He thought perhaps the unemployed are reporting more meaning in their activities, which is puzzling and worth exploring further. Analysis of the EDT data could examine the U-index per activity comparing employed with unemployed respondents. Fleche reported an unusual relationship between passive leisure and unemployment in the EDT data that merits further exploration.

Using Weights

Deaton stressed that proper use of weights for analysis of data from the DRM and DRM hybrid methods is critical. The BLS weights for ATUS—while correct—are complicated and difficult to use. Improper use of weights can result in anomalous findings. Articles that do not report exactly what weights were used and how they were used should be discounted in the literature.

Houle observed that understanding how and when to use which weights is a problem in Canada as well. The Canada data include episode weights and person weights. The episode weights should be used when the activity is the unit of analysis, yet many data users incorrectly use the person weights. Methodologists at Statistics Canada are preparing clear examples and instructions for correctly using the weights. In general, more explanation needs to be provided to better inform data users.

Biomarkers

Daly and Yamamoto are using different biomarker approaches in their work. Conal Smith wondered whether it would be useful to identify a common set of relevant biomarkers, their properties, and their relationship to the DRM. Nielsen noted that participants of a previous Roybal Center meeting on positive psychobiology commented that much of the biomarker work

²¹ See McGraw, A.P., Larsen, J.T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. *Psychological Science*, 21, 1438-1445.

has been conducted from a relatively narrow stress or disease perspective.²² A broader range of measures to capture biomarkers associated with both positive and negative affect should be considered. Liam Delaney posited that the relevant aim should be to discover the significance and meaning of the connection between affect fluctuation and biological fluctuation, rather than focusing on validation of biomarkers.

Individual versus Population Levels

Jacqui Smith noted that the strengths and weaknesses of the DRM and EMA methods are different depending on whether one is interested in individual- or population-level findings. Some of the presented work focused on variability within an individual over different days and specific aspects of behavior. At the population level it is important to examine patterns by subgroups of the heterogeneous population (e.g., unemployed, single working parent). Over a period of time the routine aspects of daily life—in conjunction with affect and biological fluctuations—can have a significant impact on health. The DRM appears suited to assessing these issues at the population level. Michelson agreed that there is a tremendous amount of untapped data available to study time use at the population level in a more detailed way (i.e., to examine subgroup differences in time use including data on where the person is and who they are with during the activity). Multitasking is another understudied phenomenon.

Methodological Developments

Collecting DRM Data

Arie Kapteyn, Dornsife Center for Economic and Social Research, University of Southern California

The ATUS is a conversation between the respondent and interviewer using CATI in which the interviewer codes the responses. Denmark collects time use data online including time interval and secondary activity. The Netherlands uses a diary app and self-administered time use surveys with smartphones. The interval at which a respondent is prompted to provide a diary entry can be adjusted. An analysis comparing the Dutch app and a traditional paper and pencil diary entry collected a year earlier demonstrated that the distribution of time across activities is similar between the two modes.

The American Life Panel (ALP) collects data on self-defined episodes.²³ The design mimics the original DRM: respondents have to define the episodes of their day, name them, provide the time frame, and answer questions about each episode including affect, location, and companions. Although there are many useful questions to ask per episode, asking all of them could lead to long and tedious surveys.

²² See <http://www.nia.nih.gov/about/events/2013/positive-psychobiology> for a description of the meeting and link to the summary report.

²³ More information about the ALP can be found at <https://mmicdata.rand.org/alp/index.php?page=main>.

There are different strategies for reducing respondent burden. ELSA preselects seven episodes, asks how long the person spent in each episode, and asks four affect questions per episode. The approach used in ELSA is to preselect seven activities: watching TV, working or volunteering, walking or exercising, engaging in health-related activities other than walking/exercising, traveling or commuting, spending time with family or friends, and spending time at home alone. The respondent indicates a duration for each of these activities and then rates strength of feeling on a scale of 0 to 6 for four affects: happy, interested, frustrated, and sad. The ALP asks questions about self-defined episodes for only part of the day. PATS and the Panel Study of Income Dynamics (PSID) supplement on Disability and Use of Time (DUST)²⁴ both randomly sample three episodes. The implications of these various iterations of DRM approaches—e.g., how many episodes or what portion of the day, self-defined episodes versus preselected activities—are unknown.

The Gallup-Healthways Wellbeing Index and HWB-12—a set of 12 questions to assess hedonic wellbeing developed by Jacqui Smith and Stone and included in the 2012 administration of the Health and Retirement Study (HRS)—use aggregate measures. Gallup includes happy, smiling, enjoyment, sad, angry, stressed, worried, and depressed, and the HWB-12 uses happy, content, enthusiastic, frustrated, sad, angry, tired, stressed, lonely, worried, bored, and pain.

Kapteyn noted that the ELSA method is an interesting approach, but little has been done to compare it to other methods used in large-scale surveys. Similarly, little is known about how approaches such as those used in ELSA, HWB-12, and Gallup compare to results obtained with a full DRM administration. Regressions of “net affect yesterday” in the HWB-12 on “net affect during the seven ELSA episodes yesterday” indicate that the seven affect episodes combine linearly. However, the lowest and highest net affect have a significant additional effect, indicating nonlinearity. These results merit further exploration.

One original impetus for developing the DRM was efficiency. Smartphones make this goal even more attainable. There is a plethora of off-the-shelf technology that could be leveraged to administer the DRM or DRM-like instruments, and these should be rigorously explored. Kapteyn noted that mode does not appear as important as determining the relationship between the variations for asking about affect during episodes yesterday. Further work is needed to understand the relationship between the DRM variations and to determine the optimal approach for the research questions of interest (e.g., how many affects, how many episodes and how selected, what type of scale for rating affect).

Single Day DRM—Suitability for Survey Work *Somnath Chatterji, World Health Organization*

The WHO conducts the longitudinal Study on global AGEing and adult health (SAGE) on a sample of about 45,000 adults aged 50 years and older from nationally representative samples

²⁴ See <http://simba.isr.umich.edu/zips/zipsupp.aspx#DUST13> for more information on this PSID supplement.

in China, Ghana, India, Mexico, Russia, and South Africa.²⁵ A short version of the SAGE questionnaire focused on health and wellbeing is included in a collaboration between SAGE and the International Network for the continuous Demographic Evaluation of Populations and Their Health (INDEPTH), which includes field sites in Bangladesh, Ghana, India, Indonesia, Kenya, South Africa, Tanzania, and Vietnam.²⁶

The SAGE survey is 2 hours long, and, therefore, adding the full DRM is not feasible. SAGE's adapted DRM divides the day into morning, afternoon, and evening. These portions of the day were used to ensure assessment of consecutive time points, rather than a random selection of episodes. Respondents are randomly allocated to morning, afternoon, evening, or full day and asked about a maximum of 10 episodes. For each episode, respondents are asked for their activity (selected from a list of 23 items chosen based on pilot work), who they were with, and a rating on a scale of 1 to 3 for the following affects: worried, rushed, irritated/angry, depressed, tense/stressed, calm/relaxed, enjoying. Wave 2 of SAGE reconstructed the activities during the whole day and randomly sampled a selection of activities about which detailed questions were asked.

WHO researchers conducted two smaller studies to compare the partial and full-day SAGE data. These studies used partial day sampling and then returned to the same respondents to do the full DRM on the same day of the week a week later. The correlation between affect reported on the same activity with the partial DRM on day 1 versus the full day DRM on day 7 was about 0.5. In a separate study, investigators used accelerometers to examine the relationship between the DRM methods and physical activity to test the hypothesis that individuals engage in fewer activities as their health declines, leading to diminished wellbeing. If this hypothesis is correct, then it is possible that improving the selection and quality of available activities could help individuals maintain wellbeing even if health status declines. Analyses were conducted to examine the differences between the full DRM and the partial version. Respondents who were asked about the aggregate day report higher positive affect and lower negative affect when compared to those who were asked to reconstruct specific activities; this finding is consistent with Schwarz' prediction that there will be different findings in assessments of experiences versus general feelings about activities.

Results from SAGE are similar to those reported in the literature on associations between affect and time use. Activities least associated with positive affect included working, subsistence farming, housework, providing care to someone, traveling/commuting, preparing food, and watching children. Activities most associated with positive affect included intimate relationships, religious activity, watching TV, playing games or cards, reading, and listening to the radio. In a separate analysis of data from China, individuals who were in the best quintile of health reported more positive affect than those in the worst quintile of health. With the

²⁵ See <http://www.who.int/healthinfo/sage/en/> for more information about SAGE.

²⁶ See <http://www.who.int/healthinfo/sage/indepth/en/> for more information about the SAGE-INDEPTH collaboration.

exception of Mexico, data from SAGE countries plus Spain demonstrate diurnal variation in positive and negative affect.

The data from the SAGE and INDEPTH studies are all in the public domain. Chatterji noted several conclusions supported by this collection of WHO studies and analyses:

- The single-day DRM produces patterns of results very similar to those reported previously with EMA.
- Short versions of the DRM produce patterns very similar to the full-day DRM.
- DRM results are reproducible with variations by day.
- There is a consistent relationship between DRM findings and health status, age, income, education, and social networks and the broader environment.

Chatterji indicated that the WHO is interested in pursuing further work to validate DRM findings with biomarkers of wellbeing and stress, and in trying to understand the relationship between personality measures and how personality mediates the relationship between health and wellbeing. He also wondered how results compare between random sampling of events and full day DRM, and the impact of asking an imbalanced number of affective questions (e.g., more positive adjectives than negative) versus an equal number of positive and negative adjectives.

The Partitioning of the Day and the Heterogeneity of Daily Experience

Dylan Smith, Stony Brook University

Stefan Schneider, University of Southern California

Differences in how respondents divide their days into episodes and the potential implications of these differences are not well understood. The full DRM instructs respondents to “think of [the] day as a continuous series of scenes or episodes in a film” and to identify the start and end time of each episode. Examples and suggestions are provided, but respondents define their own episodes. Not every respondent reports the same number of episodes: the range is 4 to 30 episodes per day.

The number of episodes reported is associated with age, education, and income. The mean number of episodes is higher for those with more education. The number of DRM episodes reported is correlated with the number of episodes reported a year later and with reports of depression, pain, and fatigue.

There are several possible explanations for the differences in number of episodes reported that require further study.

- Some groups simply may have “richer” lives with more daily episodes than others.
- Individuals may have different strategies or criteria for partitioning the day.
- Individuals vary in cognitive complexity, memory, motivation, and storytelling abilities.

These issues could be explored with cognitive interviews, by following individuals in real-time (possibly via video) to assess how they later define episodes, and/or by comparing information

yielded when different instructions are provided. Houle observed that in Canada the number of episodes reported was consistent across two modes: CATI with instructions on partitioning episodes and electronic self-administrated questionnaire without explicit instructions.

Approaches to partitioning could be assessed experimentally by asking respondents how they would create episodes from samples of daily activity logs. Comparing the structural characteristics of episodes for high- and low-episode respondents (e.g., do they have multiple episodes of the same activity, do they have more interactions with others) could reveal how respondents define their episodes and elucidate a pattern or theory that explains the variation.

Another consideration for future study is the link between episodes and hedonic experiences. It is not known how well a single rating represents the feeling during an entire episode because it cannot be assumed that hedonic experiences are constant within an episode, particularly when episodes are of varying length and quality. More work could be done to better understand how respondents select an affect rating for an episode when the underlying experience includes fluctuations. It is unclear whether respondents select an affect for an episode rating based on an average of experiences during the episode, a randomly selected time-point within the episode, the dominant mood of an episode, or a combination of the peak and end of the episode. This could be investigated by asking for affect ratings during a specific time segment of an episode or by asking for ratings during multiple time segments per episode.

Understanding how respondents report affect for an episode is important for examining dynamic within-person processes, such as diurnal cycles, and linking of emotions with real-time biological data (e.g., blood pressure, cortisol). This is currently difficult because affect ratings for an episode collected with the DRM or DRM variants are not linked to a specific point in time. Meeting participants had different opinions on the importance and implications of measuring (or not measuring) possible within-episode variation of hedonic experiences. Most agreed that it would be important to understand the extent to which within-episode variation exists, which could be assessed by comparing DRM and EMA data. Schkade suggested that these issues could also be explored with simulation modeling. Days could be constructed with episodes created based on an assumption of the distribution of the experience. Different methods could be used to determine how the aggregate results differ.

Nielsen posited that what might explain how individuals delineate episodes could also explain how a respondent applies a single affect rating to an episode. For example, people with higher levels of emotional awareness or cognitive ability might parse the episodes and affect ratings differently.

Dylan Smith and Schneider raised a related issue regarding duration weights. Longer episodes are weighted more strongly than shorter ones in duration-weighted averages. This works as long as affect reported for shorter and longer episodes represents an average of the episode (rather than a peak or end of an episode). Respondents might have a less clear recollection of longer episodes and therefore report a more generalized affect for the time frame. In other words, longer episodes might be more prone to heuristic responding.

DRM Aggregation Issues

David Schkade, University of California, San Diego

Schkade commented on several issues related to aggregation of DRM data: a focal point within a particular episode, duration weighting, aggregating affects, and time gaps. In his work, Schkade has found that respondents report an average of 1.7 to 2.3 activities per episode. In order to determine which activity likely contributed most to affect during a given episode, Schkade asked respondents to identify a focal activity within the episode: "If you checked several things you were doing at the same time, please circle the one that seemed the most important to you at the time." Random variation in results was reduced and correlations increased when this prompt was used.

Schkade noted that it has not been demonstrated whether duration weighting improves predictive value of affect measurement. In fact, it is unusual for duration-weighted results to differ from unweighted results in terms of the correlation between affect and life satisfaction or in predicting evening affect from morning affect. This could be explained if the episodes were all the same length, but another possibility is that there is noise in the duration. Schkade asserted that duration weighting is theoretically correct but in practice does not improve results compared to equal weighting of episodes.

The U-index is an informative way to derive one response per episode. Schkade tried several ways to aggregate across affects, not across episodes. Although some affects have a unique correlation pattern, it may be useful to combine negative affects because they are less correlated than positive affects. There is one positive and three negative affects in ATUS.

Gaps in time use data also present problems for aggregating and analyzing data. Respondents often leave time gaps or have overlapping episodes when using a free form diary. To address this, Schkade proposed normalizing responses such that each day is assigned a value of 1 and episode duration is represented as a proportion. This approach might result in losing information, but it would reduce the noise.

Short Versions of the DRM: Ready for Prime Time?

Jacqui Smith, Institute for Social Research, University of Michigan

Jacqui Smith has been involved in developing a short version of the DRM for the Health and Retirement Study (HRS).²⁷ A comparison of the HRS DRM measures and the full DRM is shown in Table 2. The HRS core survey takes about 2 hours per person to complete. The 2009, 2012, and 2014 HRS field periods included a brief day reconstruction instruction, general questions about the day and context, participation and time spent on selected activities, and affect ratings for each activity. Respondents were asked to rate affect on a scale of 0 to 6. Three positive and three negative affect adjectives were used, the first four of which were

²⁷ See <http://hrsonline.isr.umich.edu> for information about the HRS.

harmonized to ELSA: happy, interested, frustrated, sad, content, bored, pain. Data are available from the 2009 pilot and 2012 administrations.²⁸

This short DRM version has also been developed for internet and telephone administration, in which start/stop times were collected and the eight activities and affect ratings were randomized. A separate study of these versions tested mode effects: Research on Well-Being and Use of Time (ROBUST).²⁹

Table 2: Comparison of full DRM 2004 and HRS 2012 DRM measures

	DRM 2004	HRS 2012
Mode	Paper	Paper
Length	45+	10+
Reconstruction	Long (detailed)	Brief
Time Use	Start/Stop episodes	Activity duration
Activity Retrieval	Temporal Self defined activities	Non-temporal Targeted activities
# Activities	16	8
Affect Rating	12 Feelings during episode Fixed order; 0-6 scale	6 Feelings during activity Fixed order; 0-6 scale
Activity context	Yes	No
Day context	Yes	Yes

The HRS short version of the DRM tested well on indicators of precision. In 2012, 83 percent of respondents provided complete information on the self-administered questionnaire. Missing data were related to older age, lower education, and lower cognitive ability. The instrument was relatively reliable and yielded consistent activities over time. A 2-day study found high retrieval consistency in a comparison of day 1 end-of-day and day 2 reconstruction for activity participation (90 percent agreement), time duration, and affect. The largest discrepancies were episode start/stop details for activities such as household chores. Consistency was not associated with age or memory. Findings from the ROBUST study demonstrated precision on approximation of diurnal sequences and mixed feelings by activity and time of day.

²⁸ The 2012 questionnaire can be found at http://hrsonline.isr.umich.edu/modules/meta/2012/core/qnaire/online/HRS2012_SAQ_Final.pdf.

²⁹ More information on the ROBUST study is available at <https://community.isr.umich.edu/public/pag/ResearchProjects/LongitudinalandSurveyResearch/ROBUST.aspx>.

The HRS short version and ROBUST activity participation and activity durations are similar to findings from ATUS and PSID-DUST, and there are similar age group patterns of happy and sad by activity. Findings are discrepant for household chores and walking/exercising. It is difficult to compare affect because of the complex weights involved. The mode effects were consistent with the literature. There was a positive bias for the CATI/CAPI versions compared to web and paper versions and no mode effect was found for negative affect.

Jacqui Smith expressed interest in examining the extent to which experienced wellbeing predicts mortality, which can be tested with HRS data. More information will be available after the current HRS field period ends and the vitality statistics are updated. Krueger agreed that it will be important to examine whether experienced wellbeing is a predictor of longevity, which could suggest the possibility of using national wellbeing data to predict mortality.

Jacqui Smith suggested several central research questions that need to be addressed to understand the usefulness and appropriateness of short versions of the DRM. Kahneman and colleagues (2004) noted that detailed description of the previous day, replication of real-time experiences, accurate retrieval of specific episodes, objective circumstances of episodes, and multidimensional episode-related affect are essential components of the DRM. Consideration might be given to other essential components, and to validating the short version against the full DRM.

Discussion

Schwarz inquired whether the HRS asks for generic reports about activity types or for a report of a specific experience during the last time the person did the activity. Asking “How did you feel the last time you exercised?” might elicit different types of responses than “When did you last exercise, where were you, who were you with, how did you feel, etc.?” Krueger added that these findings will over represent respondents who do the selected activities frequently, which will affect the weights. Jacqui Smith responded that the HRS Consumption and Activities Mail Survey asked how much time is spent on each activity, which would enable some cross validation.

Carrie Exton wondered what data might be missing when the activities are preselected. Nielsen noted that the activities selected for the HRS version were based on time use data and the age of respondents; the selection of activities might differ depending on the sample. Stone countered that this approach could miss activities that are important but not frequent. Researchers need to be clear about how and why activities are preselected and what implications the selection has for interpreting results. Exton noted that this HRS short version is excellent for specific research questions, but it might not work for a national statistics organization because it misses 33 percent of the day and, therefore, will not yield an estimate of experiences of the whole day at the population level.

What constitutes the essential components of the DRM? Stone asserted that these hybrid methods are useful for certain contexts and research questions but he did not consider them “the DRM.” Krueger disagreed. Schwarz argued that the nomenclature is irrelevant; the key is to be clear about the strengths and limitations of each instrument and its variations and for

what purposes each is appropriate. Many agreed that affect without time use is not very informative, and time use is a key element of what distinguishes the DRM from the Gallup measures.

Aggregation over the Day versus Mixed Methods Estimating and Weighting of the DRM

Joseph Schwartz, Stony Brook University

Angus Deaton, Princeton University

Schwartz and Deaton spoke about issues related to sampling, weighting, testing with simulation, and other methodological issues. Deaton asserted that an advantage of the DRM is its ability to measure how much time people spend in what affective states. The original DRM collects a diary for the entire day, which enables a sample of people-minutes. ATUS samples three episodes from each respondent without regard to duration of the episodes. More person-minutes will likely be included from individuals with fewer total episodes per day than from individuals with more, shorter episodes per day because three episodes are sampled in both cases. The primary concern is precision: the standard error of the U-index will be larger for a respondent with three short episodes using this method compared to a full-day sample because change in the U-index is based on change in time use.

Schwartz advised that researchers should ensure and document that the unit of analysis for the question is the unit of analysis of the observation being used in the dataset. The utility and validity of the DRM depends on what type of question is being asked and might vary by question. Schwartz recalled that Schkade is interested in the average affect associated with an activity and is not interested in people or episodes; for this inquiry, Schwartz recommended that all data are aggregated by activity. Conversely, Daly is interested in a personality measure of self-control and its relationship with affect for multiple episodes and health indicators. To do that type of analysis, the average affect for the individual needs to be estimated, and its accuracy at the person-level is critical or else the analysis will attenuate the relationship with self-control.

Schwartz expressed concern and skepticism about the ability of respondents to accurately recall affect from the previous day without reinstatiation. Recall is better with full reinstatiation, but even then it is a challenging task. It is possible and maybe even likely that the measure is capturing respondents' estimates of how they think they usually feel during an activity or episode. He asserted that we need to better understand the within-person variability of affect reports as he fears that the correlation is, in actuality, very low.

Do We Know If Reinstatiation Is Important?

Norbert Schwarz, University of Southern California

The DRM relies on recall, yet respondents' memory is poorer than is assumed. Responses are typically based on a mix of general knowledge and recall of actual events. These types of reports are not *wrong*, but they do not accurately capture specific experiences. Through

experiments on autobiographical memory, Bradburn and colleagues showed that “respondents will use any information they have in order to generate a reasonable answer.”³⁰ Respondents draw on intuitions that the researcher and audience share and therefore findings appear congruent and logical. However, there are systematic biases.

Research has demonstrated that respondents use the present as a benchmark to reconstruct past behavior and experience.³¹ If there is no reason for the respondent to believe the past was different from the present, then responses are actually based on the present. If there is reason to believe the past was different, then the respondent adjusts his or her response for the perceived change. This mental process overestimates stability in many domains, such as income, alcohol or drug consumption, pain, and media use. For example, a longitudinal study of pain using daily reports and recall after 1 week demonstrated that last week’s pain is reported to be more similar than today’s pain than warranted. Respondents also tend to overestimate the successes of medical interventions, retrospectively reporting greater baseline pain after receiving a treatment.

Retrospective reporting of feelings is particularly problematic. Feelings are fleeting and not well represented in memory. Individuals can introspect on feelings, but require episodic knowledge and semantic knowledge to reconstruct the feeling later. EMA measures feelings concurrently and the response in the moment can be based on introspection (i.e., how do you feel right now?). Episodic reports can capture past feelings with some accuracy, provided the episode is recent and is relived in detail. EMA cannot be a gold standard in all situations, because a lot of activities are inaccessible by the method (e.g., during driving, during medical procedures, during important meetings, etc.). At times, the DRM can be better than EMA because it attempts to capture the most recent episode and guide the respondent through reliving the episode to recall the feeling. This type of reinstatement is currently the best approach for retrospectively measuring feelings during specific episodes.

Global reports characterize how respondents generally feel during a type of activity. The respondent is not encouraged to relive the episode. Instead, the respondent focuses on the key attributes of the activity or episode (i.e., a focusing illusion) and reports general beliefs about how he or she usually feels during that activity. Global reports are based on different inputs and overestimate intensity compared to concurrent and episodic reports. Predictions are based on the same inputs as global memories: respondents focus on the main characteristic of the activity in the absence of experience. Predictions, choices, and global memories are all based on the same intuition and provide consistent and reliable data. Ironically, the less respondents rely on actual experiences, the more consistent data they provide. Future behavior is predicted by memory, not actual experience, which may be very different.

³⁰ Bradburn, N.M., Rips, L.J., & Shevell, S.K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, 236, 157-161. Retrieved February 10, 2015 from <http://shevell.uchicago.edu/publications/Bradburn%20Rips%20Shevell%20Science%201987.pdf>

³¹ Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, 96, 341-357.

The most appropriate type of measure depends on the goal. Global measures are appropriate for predicting behavior. Concurrent methods are ideal for understanding experiences. Episodic methods are a reasonable compromise when concurrent data collection is not feasible. Table 3 provides a list of types of measures and the level of reconstruction used.

Table 3: Levels of reconstruction used in different measures

Measure	Level of Reinstatiation
EMA	Concurrent
DRM	Recent episodes with reinstatiation
ATUS	Recent episodes without reinstatiation of details
Yesterday	Recent episodes without any reinstatiation
Event Reconstruction Method (ERM)	More distant episodes, reinstatiation of episode with limited context
Usually	General knowledge
Future	General knowledge

Stone noted that all research questions are not typically known a priori for large-scale national surveys. National statistics organizations will need guidance about which measures to include even without knowing the exact research questions that will be asked of the data.

Daly suggested a research priority is to systematically define grades of reinstatiation, randomize them, and compare results among the methods. One problem is that even if the instrument is randomized, it is unclear that the episodes among respondents in a sample are comparable, making it difficult to determine whether observed differences are due to the methodology. An ideal experiment would involve controlling events for the participants.

Michelson reiterated that one of the values of the full-day time use diary is that it prompts respondents to enumerate activities in chronological order, which makes the day clearer and potentially more accurately reported. On the other hand, the accuracy depends on how well the questions prompt the respondent to reinstatiate each episode.

EMA versus ATUS Wellbeing

Alan Krueger, Princeton University

Krueger presented methods, procedures, and findings from a study comparing EMA and a telephone recall survey that approximated the ATUS wellbeing module. A convenience sample of 168 employed individuals carried a handheld device that asked questions about their activities in real-time for 3 consecutive days (Thursday, Friday, Saturday). The data were collected between April and August of 2008. The day was divided into six even intervals, and the handheld device beeped randomly within intervals, leading to approximately 3,000 awake moments collected across the sample. Gallup administered a telephone recall survey (P-ATUS) to the same individuals on each day following EMA data collection (N=478 interviews). The goal of the study was to compare P-ATUS to EMA in aggregate and in 105 overlapping moments.

The handheld device used for the EMA data collection recorded duration of the activity and asked for the following details: location, activity (multiple responses allowed), whom with, feelings of happy, angry, tired, stress, interested, and pain on a 0-6 scale, whether the person had taken medication since the last prompt, and whether the person had caffeine, cigarettes, or alcohol in the last 30 minutes. An Actiheart device continuously measured physical activity, caloric expenditure, and heart rate. Cortisol was measured six times a day after each EMA prompt.

Gallup modified the ATUS Blaise to create the P-ATUS. After completing a diary for the whole day, the SWB module randomly selected three awake episodes and asked for affect ratings. This was actually a mistake—Gallup was supposed to ask for affect ratings on all the episodes.

Krueger reported several results of the study. Comparison of location reported in EMA and P-ATUS with a sample of 2,396 overlapping moments resulted in a 73.7 percent exact match, 79 percent match within 15 minutes, and 20.8 percent mismatch. EMA data demonstrate that the respondents were multitasking 26 percent of the time. On average, 1.32 activities were being done at any given moment. Comparison of activities in EMA and P-ATUS with a sample of 2,360 overlapping moments resulted in a 62.5 percent exact match, 71 percent match within 15 minutes, and 29.2 percent mismatch. The greatest area of mismatch, both for location and activity, was reports of being at work (location) or doing work (activity). Reports of work were higher with P-ATUS than with EMA.

A subject-level analysis demonstrated that some respondents are atypically bad at matching their reports within 15 minutes. It would be interesting to characterize these types of people.

Differences in the average of subjects' feelings ratings between the two methods were statistically significant for all affects except for happy. The correlation between affect from EMA and P-ATUS at the person level ranged from 0.74 for interested to 0.86 for pain. Adjusting for sampling variance in the person averages increases correlations to as high as 0.94 for pain.

For the 105 moments sampled in both EMA and P-ATUS, the correlations between the two methods ranged from 0.41 for happiness to 0.54 for pain. The correlation of the U-index measured in overlapping moments was 0.54. These correlations, while lower than might be expected, are still nontrivial. They could be skewed because the P-ATUS refers to an episode and the EMA data are for a moment in time. A comparison of the U-index by activity in EMA versus P-ATUS yields a 0.9 correlation if weighted by number of observations.

Krueger noted several conclusions. Data indicate a reasonable match rate between real-time reporting and P-ATUS—better for location than activity. Discrepancies might be due to mistakes in EMA or P-ATUS. Work hours are overstated in P-ATUS. Less precise or longer recall questions about work elicit responses that blur work and nonwork activities. Nonresponse in EMA (skipped beeps) can be studied. The P-ATUS module adequately characterizes individuals' days and activities compared with real-time reporting. Overlapping moments are less highly correlated. The instruments could be validated further with physiological data. ATUS activities

appear to reflect meaningful variation in heart rate. These data could be used to create an American Heart Rate Index.

Conal Smith suggested that if the DRM is performing as expected, then the correlations should improve with more EMA moments. Stone offered that 12 to 18 sampled moments per day in EMA can yield a reasonable compliance rate. Many questions could be answered with denser EMA data and a subsequent DRM.

Measuring the Effect of Thoughts on Daily Wellbeing

Talya Miron-Shatz, CureMyWay and Ono Academic College, Israel

Miron-Shatz's hypothesis is that thoughts add to the prediction of wellbeing. Shimmack and colleagues evaluated the impact of thoughts about important life domains on life satisfaction judgments using open-ended and closed-format retrospective reporting.³² They showed that thoughts are stable over time and that certain domains (e.g., family relations, academic performance) are thought about more frequently than others (e.g., financial satisfaction). These findings suggest that frequency and valence of thoughts are important.

There are several challenges to measuring thoughts: (1) when to measure: retrospective (DRM or ERM) or during (EMA); (2) how to measure: closed or open set of thoughts; and (3) what to measure: frequency, valence, intensity of thoughts or some combination thereof. Miron-Shatz presented methods and results for two studies she conducted using ERM and DRM on thoughts about financial security³³ and a third ERM study on a variety of thoughts associated with major life domains.³⁴ Results from the first two studies suggested that (1) the consideration of financial security was as important to participants' life as their monetary assets and that (2) women who mentioned financial concerns in their thoughts about the future (e.g., retirement, college tuition payments, making ends meet) reported lower life satisfaction compared with women who did not raise such concerns. A conscientiousness personality variable predicted thoughts about the future, whereas personal or household income did not.

The third study demonstrated that the emotional valence of thoughts about major life domains was an important predictor of current and prior life satisfaction. Greater emotional valence was often associated with the domains thought about most frequently. These findings suggest that thoughts are a relatively stable aspect of wellbeing and require further investigation, or even inclusion in time use surveys.

³² Shimmack, U., Diener, E., & Oishi, S. (2002). Life-satisfaction is a momentary judgment and a stable personality characteristic: The use of chronically accessible and stable sources. *Journal of Personality, 70*, 345-384.

³³ Miron-Shatz, T. (2009). Am I going to be happy and financially stable? How American women feel when they think about financial security. *Judgment and Decision Making, 4*, 102-112.

³⁴ Miron-Shatz, T., Diener, E., Moore, T., & Saphire-Bernstein, S. (2013). Charting the internal landscape: Affect associated with thoughts about major life domains explains life satisfaction. *Judgment and Decision Making, 8*, 603-616.

Stone asked how thoughts were measured. Miron-Shatz explained that respondents were asked how often they thought about each of the given set of items. Responses ranged from never, rarely, at least once, to many times a day. Respondents were then asked how they feel when they have that thought. Daly noted that it is difficult to assess thoughts in real-time because asking the question distracts the person from the thought. Schwarz observed that these data do not indicate when individuals think about the episode versus when they are thinking about something else. Miron-Shatz noted that episodes are, by definition, at least 20 minutes long whereas thoughts are fleeting yet can still be impactful; thoughts and episodes do not necessarily overlap.

Day Reconstruction and Linkage to Other Data

Liam Delaney, Behavioural Science Centre, University of Stirling

Delaney and colleagues have been using variations of the DRM in their research for the past 10 years and see potential to link DRM data to real-time biological data, geographical information systems, and decision-making parameters.

One of Delaney's studies integrated biological assessments with the DRM and found that negative affect was significantly related to elevated heart rate, suggesting that negative affect is a good measure of potential health strain. Tiredness was associated with a substantial decrease in heart rate, and life satisfaction did not predict heart rate. Low morning cortisol appeared to be linked to a steep rise in positive affect from morning until evening, particularly amongst the distressed. The biological findings support the assertion that DRMs assess emotional experience and not life evaluation because each has distinct correlates.

In a related line of work, Delaney is using the DRM to analyze diurnal patterns of desires and self-control. The adjectives used in the DRM for this study were tailored to match research questions about decision making and self-control failures (e.g., frustrated/annoyed, stressed, tired, hungry, drunk, focused, thirsty). This study is soliciting respondent feedback on what they desire in the moment and then determining whether they resist or act on the desire. Reports of feeling in control and being focused decrease throughout the day and tiredness increases.

Delaney and colleagues are conducting pilot studies using geographical information systems to map health and emotion onto geographical data that can be aggregated to identify patterns of heart disease and life satisfaction. For example, stressful traffic patterns can be mapped.

Delaney reflected that the DRM is a flexible framework that allows examination of diurnal patterns and is not a fixed instrument, echoing Stone's earlier statement introducing the DRM. Abridged versions of the DRM can serve different purposes. Wearable bio-measurement devices can be combined with the DRM in some situations, but are likely not feasible for large population studies. The DRM enables researchers to examine real-world decision making in a way that is not possible in a laboratory context. Experienced wellbeing and life satisfaction measures can yield contradictory outcomes because they are measuring different constructs. Reconciling these differences and clearly identifying measures that are appropriate for evaluating policies is important.

Integrated Assessment Framework DEDIPAC

Thomas Kubiak, Johannes Gutenberg University Mainz

The main objective of the Determinants of Diet and Physical Activity (DEDIPAC) consortium is to understand the determinants of diet, physical activity, and sedentary behaviors and to translate this knowledge into effective promotion of a healthful lifestyle.³⁵ As part of DEDIPAC, an integrated assessment framework will be developed that uses both the DRM and EMA.

The DEDIPAC data collection methods include a 24-hour dietary recall, a food record collected via EMA, and physical activity measured by an accelerometer. The DRM, EMA, and global positioning system mapping are used to measure within-subjects psychosocial and behavioral determinants. Between-subjects determinants are collected with trait questionnaires and demographics. The accuracy of dietary recall is improved by being framed within the DRM. Bar code scanning on food provides objective data on food intake. The accelerometer is designed to detect specific kinds of activities such as sitting, walking, or cycling.

DEDIPAC researchers are planning to launch a pilot study focused on trait loneliness, wellbeing, social norms, social setting, and exposure in the Netherlands, Germany, and Norway. It will be the first test of the assessment framework and will use both the DRM and EMA.

As Delaney noted, accelerometry and geographical mapping combined with DRM data present new opportunities that could lead to valuable insights. One challenge, particularly when the research questions focus on diet and eating, is that the data collection makes users more aware of their dietary intake and can change their eating behavior.

Discussion of Methodological Developments**Communicating Survey Benefits to Respondents**

Houle noted that communicating the benefits and purpose of surveys to respondents is critical for maximizing participation, especially when these methods are used as part of a long interview. Conal Smith commented that presenting examples of findings or analysis from other studies to demonstrate the possible uses and benefits of the data could be persuasive. Schwarz suggested that framing the survey in terms of wanting to learn about how people live their lives and experience their days is more appealing to respondents than a technical academic purpose. Nielsen noted that the United Kingdom Office of National Statistics framed its interest in SWB data as Schwarz described and conducted public outreach before using the questions. Veenhoven suggested telling respondents that others can learn from their experiences.

Multiple or Secondary Activities

Krueger noted that little research has explored secondary activities and how they relate to affect (e.g., listening to music versus listening to music while commuting). ATUS does not collect data on secondary activities or multitasking. Schkade observed that secondary activities

³⁵ See <https://www.dedipac.eu> for more information.

could be treated in several ways. When people cite more than one activity at the same time, it needs to be determined whether that should be treated as its own unique cluster, as two different activities with separate influence, or both. Perhaps more can be learned by thinking about how activities fit together instead of looking at them in isolation.

Reflecting on the DEDIPAC framework, Krueger suggested that collecting data on secondary activities would be important for studies focused on dietary intake. Comparisons of time spent eating by French and American women appeared to show that American women eat less. However, American women were more likely to eat as a secondary activity and when these data were considered, they showed that American women actually ate more than their French counterparts.

Monitoring Effects

Schwarz cautioned that every measurement is also a treatment, and he expressed pessimism that the item wording could sufficiently address monitoring effects. Monitoring effects are unavoidable, and researchers need to be aware of and account for them.

Trade-offs

A recurring theme throughout the conference was the need to study, document, and assess the trade-offs between different approaches to obtaining time use and affect data with the DRM and related methods in order to inform decisions about the best approach for a given purpose.

Schwarz observed that the full DRM uses reinstatiation to activate thoughts and feelings about the previous day's activities. The ATUS module selects three random episodes from the previous day's time use data; however, no activity is reinstatiated in detail. The trade-off between respondent burden and reinstatiation leads to differences in how findings can be characterized—is the instrument measuring specific remembered experiences as with the full DRM, or a generalized idea of feeling? Schwarz predicted that methods that generalize have more intuitively appealing, consistent data. Methods that use reinstatiation will yield more accurate, yet less consistent data. Stone added that true reinstatiation will be sensitive to variation in daily life, which means these measures will appear weaker psychometrically when compared to measures of global life satisfaction. Kapteyn suggested that more work needs to be done to determine whether these approaches are truly measuring different parameters, and if so, what are the differences and their implications for findings.

Miron-Shatz suggested that asking respondents to report on the best and worst moments from the previous day is a feasible approach to capture the respondent's perspective on what is important in his or her life.³⁶ Others asserted that asking for best/worst incidents during a day yield global evaluation judgments rather than momentary affect. The best/worst moment approach does not employ any time use and, therefore, will not provide any data about the rest of the day. Jacqui Smith suggested respondents could complete a time use diary for the whole

³⁶ Miron-Shatz, T. (2009). Evaluating multiepisode events: Boundary conditions for the peak-end rule. *Emotion, 9*, 206-213.

day and then be instructed to select the best/worst episode. Stone noted that this approach differs from the intent of the DRM: it adds a judgment factor of what the respondent values.

Veenhoven indicated that one of the main advantages of the full DRM is that it induces a reinstatement of the whole day, which is needed in order to discern relative differences in the quality of activities versus the absolute rating of a given activity. For example, some people are trait unhappy and will rate their work low, but in order to assess the quality of work, an understanding of the difference between work and private activities—rather than the absolute rating of work—is needed. Michelson observed that this is an issue of unit of analysis. If the goal is to learn about subjective valance of activities, then episodes can be sampled. However, the full day is needed if one wants to learn about the structure of a day and how different people experience it.

Conal Smith summarized that the purpose of the data collection needs to be clear and the data need to be appropriate for the purpose. One episode per person is sufficient if the purpose is to obtain an aggregate measure of society as a whole. Understanding variation in mood and how it is driven by activity versus person-specific factors requires more than one observation per person. Although the ATUS may not provide the same information as the DRM, it can provide insight to topics such as risks, unemployment, subgroups' changing participation in the workforce (e.g., women), and commuting.

The Role of the DRM in Informing Policy

Angus Deaton and Alan Krueger, Princeton University

Both Deaton and Krueger remarked on the widespread influence of the 2009 Stiglitz Commission Report in bringing attention to the topic of wellbeing and encouraging governments to measure it. Wellbeing offers an alternative target for societal improvement compared to conventional economic measures, such as gross domestic product, that poorly account for non-monetary factors (e.g., leisure time).

Deaton observed that although the DRM is important and useful, it is not the answer to everything. Krueger noted that the DRM is useful because it marries time use (objective) and affect (subjective) to provide a richer, more policy-relevant picture. We are far from understanding the behavioral responses that a particular policy elicits. SWB can sensibly inform public policy without having an encompassing measure of utility. Krueger was persuaded by Kahneman that measuring misery and influencing policy to alleviate misery is important.

Krueger noted that his work in designing macroeconomic policy proposals to solve specific problems was not overly influenced by SWB considerations. In a day-to-day sense, policymakers are typically concerned with cost-benefit analysis for specific policy programs rather than large-scale macroeconomic issues. One challenging area to explore is how to account for non-monetary aspects of regulations. Time could be used instead of cost in a cost-benefit analysis to provide an alternate picture of the way we value and influence the quality of life. Conal Smith agreed that cost-benefit analysis of policies, rather than macroeconomic policymaking, is widespread among OECD governments. It is unclear how the DRM could be used in cost-benefit

analysis. Perhaps the DRM would be more informative in program evaluation efforts than in cost-benefit analysis.

Future Research Agenda Considerations

Informed by the presentations and discussions, Conal Smith offered three categories for future research: (1) filling the knowledge gaps to facilitate further development of the DRM and its hybrid approaches; (2) identifying the core methodological issues that need to be resolved in order to inform and advise national statistics organizations that are considering including SWB measures in large-scale, nationally representative time use surveys; and (3) determining research questions that would be appropriate and useful for national statistics organizations conducting large-scale experimental work.

The DRM is valuable for a host of interdisciplinary scientific purposes that aim to alleviate misery or increase happiness. The DRM is particularly well suited for evaluating programs or smaller-scale policies, such as clinical interventions or end-of-life decision making. In order to best serve these purposes, future research is needed to further develop the DRM and understand the implications of and uses for its abbreviated variants.

There are several arguments for national statistical organizations to collect both experienced and evaluative wellbeing. First, based on an exhaustive review of the literature, one of the primary conclusions of the National Research Council Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework was as follows:

To make well-informed policy decisions, data are needed on both [experienced] and evaluative wellbeing. Considering only one or the other could lead to a distorted conception of the relationship between SWB and the issues it is capable of informing, a truncated basis for predicting peoples' behavior and choices, and ultimately compromised policy prescriptions.³⁷

Second, the U-index, while not a direct measure of utility, might be a satisfactory indicator of societal wellbeing. Time use and experienced wellbeing data are needed to derive the U-index. Finally, future national experiments are unknown. Ongoing collection of wellbeing data will enable national statistics organizations to conduct research that cannot yet be predicted.

The following list summarizes the methodological issues and suggestions for future research raised by conference participants.

³⁷ National Research Council. (2013). *Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*. Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework. A.A. Stone and C. Mackie, Editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, p. 93.

Improve Ability to Inform Policy and Advise National Statistics Organizations

- Identify the key features needed in a DRM hybrid method appropriate for national statistics organizations.
- Determine the purposes and circumstances for which it is appropriate for a national statistics organization to conduct a full DRM on a subsample versus a shorter hybrid DRM on a nationally representative sample.
- Conduct research to determine how many and which affect dimensions to measure in national studies.
- Evaluate the best way to aggregate time use and affect data (i.e., the U-index) to inform policy.

Create Better Documentation

- Create a simple yet detailed guide or handbook that describes how DRM data can and should be used, with particular attention to weighting methods.
- Create a complete methodological manual that includes comparisons of DRM and DRM hybrid methods.

Expand Research in New or Understudied Areas

- Consider a broad range of biomarkers and their relationships to the DRM.
- Analyze time use and affect data of understudied subgroups.
- Create better measures of or opportunities to collect data on multitasking or secondary activities.
- Use HRS data to examine the extent to which experienced wellbeing predicts longevity.
- Conduct cognitive interviews, experimental studies, focus groups, and secondary data analyses to examine the differences in how respondents delineate episodes in their day and how these vary by subgroup, instrument instructions, or other factors and the implications of these differences.
- Determine whether alternative metrics, such as the U-index, can be used to derive new metrics from DRM-type data to provide insights into various outcomes of interest (e.g., pain).

Systematically Assess Trade-Offs

- Systematically assess how DRM hybrid approaches, such as ATUS and the HRS wellbeing module, work for a range of measures and whether the standard errors are acceptable. This could possibly be done with a simulation study.
- Conduct systematic comparisons of the EMA and DRM measures for reliability, validity, accuracy, etc.
- Empirically test and compare multiple DRM versions using various levels of reinstatement and EMA.
- Define the essential components of the DRM.
- Assess how each trade-off in the DRM-like methods compares to the full DRM and identify the implications for the types of research questions that can be addressed with each method.

- Test the full DRM and one or more DRM hybrid methods (and possibly EMA) in the same population to study the relationships and trade-offs. There might be opportunities to add onto existing studies to make these comparisons.
- Define a set of criteria by which DRM hybrid methods should be measured to determine how accurate, effective, and useful the method is for what purposes and populations.
- Determine whether a single rating of affect per episode is adequate to accurately measure experienced affect and document what is known about how a respondent chooses a rating (e.g., dominant mood, an average).

Summary of DRM Concepts

Numerous variables should be tested to understand the following questions:

- What implications does each of these choices have for the results?
- Which of these variables is required to consider it a DRM method?
- Which variables are suitable for what purposes?
- What package or possible combinations of packages are appropriate to recommend to national statistics organizations?

List of Concepts

1. Length of administration
 - a. Minimize respondent burden
 - b. Maximize reinstatement of full day
2. Mode
 - a. Self-administration
 - i. Internet-based electronic questionnaire
 - ii. Smartphone app
 - iii. Paper and pencil
 - b. Face-to-face interview
 - c. CATI
 - d. Computer-assisted personal interviewing (CAPI)
3. One day versus multiple days
 - a. Differences between weekday and weekend days?
 - b. Differences between seasons?
 - c. Is one day enough?
4. Time use of whole day or partial day
 - a. If whole day, then how documented:
 - i. Diary method from wakeup time to bedtime, respondent records start time and duration of each episode to account for whole day
 - ii. Respondent chunks day by episodes—prompted to eliminate gaps or overlaps or free form?
 - iii. How are days partitioned? What are the implications of different approaches?
 - b. If partial day, then how selected and is it attached to time use?
 - i. By time, by chunk, by activity, by episode

- ii. Random selection, preselection of activities, respondent selection
 - iii. Quantity of selection (3 episodes, 1 activity, etc.)
 - c. Can activities and/or episodes be allocated to time?
 - d. Can you do duration-weighted analyses?
- 5. Reinstantiation—what criteria are needed to determine whether reinstantiation is sufficient?
 - a. Is reinstantiation required for the method to be considered DRM?
 - b. Characteristics of a method that measures remembered experience
 - c. Characteristics of a method that measures generalized idea of feeling about a type of activity
 - d. Characteristics of a hybrid method (and why, when, and for whom would it be appropriate to use)
 - e. Is there enough reinstantiation to obtain experienced wellbeing versus generic wellbeing reports?
- 6. Questions asked about each episode—how much is needed?
 - a. Selection of episodes about which to answer questions
 - i. By time, by chunk, by activity, by episode
 - ii. Random selection, preselection of activities, respondent selection
 - iii. Quantity of selection (3 episodes, 1 activity, etc.)
 - b. Affect
 - i. How many affects needed?
 - ii. Which ones? Balance of positive and negative?
 - iii. What type of rating scale?
 - c. Who with
 - d. Location
 - e. Secondary activities
 - f. Other?
- 7. Purpose of administering the instrument
 - a. Characterizing a population
 - b. National comparisons
 - c. Determining value of nonmarket activities
 - d. Identifying changes in people's lives over time
 - e. Assessing intra-individual variability
 - f. Determining relationship between affect fluctuations and biomarker fluctuations
 - g. Making conclusions about the experiences of subgroups
 - h. Characterizing subjective valance of activities
- 8. Periodicity
 - a. Longer full-day survey less frequently
 - b. Shorter partial-day survey more frequently

CONFERENCE AGENDA

Thursday, January 8, 2015

- 8:45 am **WELCOME AND OVERVIEW**
 Michael Quick, Interim Provost and Senior VP for Academic Affairs, USC
 Arthur Stone, USC, Conference Chair
 Conal Smith, OECD, Conference Co-Chair
- 9:05 am **1. PERSPECTIVES FROM THE DEVELOPMENT TEAM**
 The team that developed the DRM will reflect on the uptake of the method, provide a summary of current DRM usage, and present perspectives on how and why the DRM was a significant achievement.
 Alan Krueger
 Norbert Schwarz
 David Schkade
 Arthur Stone
 Angus Deaton
- 10:00 am **2. SUBSTANTIVE DEVELOPMENTS I**
 2a. DRM in the American Time Use Survey
 Alan Krueger
 2b. Experienced wellbeing in France
 Sarah Fleche
 2c. Time use and the DRM in Canada
 Patricia Houle
 2d. Engagement with activities
 Ruut Veenhoven
 Discussion
- 11:10 am **BREAK**
- 11:30 am **3. SUBSTANTIVE DEVELOPMENTS II**
 3a. Fatigue assessed with the DRM
 Yoshi Yamamoto
 3b. Affect in time use surveys
 William Michelson
 3c. DRM and self-control
 Michael Daly
 Discussion
- 12:30 pm **LUNCH**
- 2:00 pm **4. METHODOLOGICAL DEVELOPMENTS I**
 4a. Variations in DRM data collection modalities
 Arie Kapteyn
 4b. Is a single-day DRM reasonable for survey work?
 Somnath Chatterji
- 3:30 pm **BREAK**
- 4:00 pm **5. METHODOLOGICAL DEVELOPMENTS II**
 5a. Do we understand how the partitioning of days works and within-episode experiences homogeneous?
 Dylan Smith
 Stefan Schneider
 5b. Presentation of DRM data—aggregating over the day
 David Schkade
- 5:30 pm **ADJOURN**

Friday, January 9, 2015

9:00 am	6. METHODOLOGICAL DEVELOPMENTS III	
	6a. Short versions of the DRM—Ready for primetime?	Jacqui Smith
	6b. Aggregation over the day versus mixed methods estimation and weighting of the DRM	Joe Schwartz Angus Deaton
10:30 am	BREAK	
11:00 am	7. METHODOLOGICAL DEVELOPMENTS IV	
	7a. Do we know whether reinstantiation is important?	Norbert Schwarz
	7b. The DRM versus EMA	Alan Krueger Talya Miron-Shatz
12:30 pm	LUNCH	
1:50 pm	8. METHODOLOGICAL DEVELOPMENTS V	
	8a. Linking the DRM to other data	Liam Delaney Thomas Kubiak
2:35 pm	BREAK	
3:20 pm	9. MOVING INTO THE FUTURE	
	9a. The DRM informing policy	Angus Deaton Alan Krueger
	9b. Future directions for the DRM	All participants
4:00 pm	ADJOURN	

PARTICIPANT LIST

Conference Organizers

Arthur A. Stone, Ph.D., chair
University of Southern California
arthuras@usc.edu

Conal Smith, Ph.D., co-chair
Organisation for Economic Co-operation
and Development
conal.smith@oecd.org

Invited Participants

Somnath Chatterji, Ph.D.
World Health Organization
chatterjis@who.int

Alan Krueger, Ph.D.
Princeton University
akrueger@Princeton.edu

Michael Daly, Ph.D.
University of Stirling
michael.daly@stir.ac.uk

Thomas Kubiak, Ph.D.
Johannes Gutenberg University
kubiak@uni-mainz.de

Angus Deaton, Ph.D.
Princeton University
deaton@Princeton.edu

William Michelson, Ph.D.
University of Toronto
william.michelson@utoronto.ca

Liam Delaney, Ph.D.
University of Stirling
liamdelaney2011@gmail.com

Talya Miron-Shatz, Ph.D.
CureMyWay and
Ono Academic College
talya@curemyway.com

Carrie Exton, Ph.D.
Organisation for Economic Co-operation
and Development
Carrie.Exton@oecd.org

Lis Nielsen, Ph.D.
National Institute on Aging
nielsen@li@nia.nih.gov

Sarah Fleche, Ph.D.
London School of Economics
s.fleche@lse.ac.uk

David Schkade, Ph.D.
University of California at San Diego
schkade@ucsd.edu

Patricia Houle, Ph.D.
Statistics Canada
Patricia.Houle@statcan.gc.ca

Stefan Schneider, Ph.D.
University of Southern California
schneids@usc.edu

Arie Kapteyn, Ph.D.
University of Southern California
kapteyn@dornsife.usc.edu

Joseph Schwartz, Ph.D.
Stony Brook University
joseph.schwartz@stonybrook.edu

Norbert Schwarz, Ph.D.

University of Southern California
norberts@usc.edu

Dylan Smith, Ph.D.

Stony Brook University
Dylan.M.Smith@stonybrook.edu

Jacqui Smith, Ph.D.

University of Michigan
smitjacq@umich.edu

Ruut Veenhoven, Ph.D.

Erasmus University Rotterdam
veenhoven@fsw.eur.nl

Yoshi Yamamoto, Ph.D.

University of Tokyo
yamamoto@p.u-tokyo.ac.jp

Observers

Michelle Boston

University of Southern California
msboston@dornsife.usc.edu

Nirmala Naidoo

World Health Organization
naidoo@who.int

Doerte Junghaenel, Ph.D.

University of Southern California
junghaen@usc.edu

Toru Nakamura, Ph.D.

University of Tokyo
nakamura@p.u-tokyo.ac.jp

Jinhyuk Kim, Ph.D.

University of Tokyo
grefur@p.u-tokyo.ac.jp

David Newman

University of Southern California
davidnew@usc.edu

Donna Metz, M.F.A., Ph.D.

University of Southern California
dmetz@usc.edu

Daphna Oyserman, Ph.D.

University of Southern California
oyserman@usc.edu

Meeting Staff

Rose Maria Li, M.B.A., Ph.D.

Project Manager
rose.li@roseliassociates.com

Maribel Estrada

Assistant, Center for Self-Report Science
maribele@usc.edu

Chandra Keller-Allen, M.P.P., Ed.D.

Science Writer
chandra.keller-allen@roseliassociates.com

Tania Gutsche

Director, USC Center for Economic & Social
Research
tgutsche@dornsife.usc.edu

Elisabeth Krebs, M.S.

Meeting Planner
elisabeth.krebs@roseliassociates.com