Memory Bookmarking:
Using *In Situ* Information to Promote Recall in Online Data Collection

by

Huiying Yan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2020

Doctoral Committee:

Professor Frederick Conrad, Chair
Professor Susan Murphy, Harvard University
Professor Norbert Schwarz, University of Southern California
Research Professor Emeritus, Roger Tourangeau
Assistant Research Professor Ting Yan, University of Maryland

Huiying Yan

yanhuier@umich.edu

ORCID iD: 0000-0002-7036-7196

© Huiying Yan 2020

Dedication

To my dearest mom

獻給我最愛的媽媽

Acknowledgements

I am more than grateful from the bottom of my heart to my adviser and dissertation chair, Professor Frederick Conrad. He has been supporting in many ways to let me pursue a bold idea for my dissertation. Without his encouragement, this dissertation would not have begun. I truly treasure all the time that he spent on our research discussions, his constructive feedback on my research study and writing, and also his emotional support for all my tough times along this bumpy journey. I am also very grateful to my other committee members - Professor Susan Murphy, Professor Norbert Schwarz, Professor Roger Tourangeau and Professor Ting Yan. I appreciate all the 1-on-1 in-person discussions as well as all of their feedback which made my research studies better and more rigorous.

I would like to thank all the financial support I got from the Daniel Katz Dissertation Fellowship, the National Science Foundation Doctoral Dissertation Improvement Grant (#1659776), the Rackham Summer Award, the Rackham Graduate Student Research Grant, the Riecker Graduate student Research Grant, the One-Term Dissertation Fellowship from the Michigan Program Survey Methodology (MPSM). These fellowships and grants enabled me to turn my ideas from paper to real data collection system and experiments. I also appreciate Dr. John Gonzales for kindly and generously offering me a Research Assistant job at Rackham in 2016 as well as the additional financial support which helped me survive the toughest time during my medical leave of absence in 2017.

I would like to thank Lucas Vickers, An Yang, and Cecile Roca for building the system with me. Without them, I would not have the data collection system I need for my experiment.

I would like to thank all my fellow students, colleagues and friends for their support, love and precious time we have spent together.

Finally, I would like to thank my family, including my husband Hung Nien, my parents and my grandparents. They have been giving me the power to face and take all the challenges in the past and those yet to come.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables vii

List of Figures ix

Abstract xi

Chapter 1 Introduction 1

References 7

Chapter 2 Conceptual Framework and Preliminary Data 10

2.1 Data Collection Framework 10

2.2 Focus of This Dissertation 13

2.3 Pilot Study 15

References 30

Appendices 31

Chapter 3 Using Text, Photo, and Map Cues via Smartphone to Promote Recall 46

3.1 Introduction 46

3.2 This Study 47

3.3 Methods 49

3.4 Results 53

3.5 Discussion 61

3.6 Limitations 64

References 65

Appendices 67
Chapter 4 Using Text, Photos, and Maps as Recall Cues in a Time Diary on the Web 81
4.1 Introduction 81
4.2 This Study 82
4.3 Methods 83
4.4 Results 85
4.5 Discussion 94
4.6 Limitations 97
References 98
Appendices 100
Chapter 5 Conclusions 110
References 113

List of Tables

Table 2.1 Experimental Conditions of the Pilot Study 15

Table 2.2 Completion Rate by Experimental Condition 16

Table 2.3 Number of Completes per Condition of the Main Study 28

Table 2.4 Demographics of Participants in the Five Experimental Conditions and Chi-square Tests for Differences Across Conditions (Main Study) 29

Table 3.1 Measures of Data Quality and Effort 51

Table A3.1 ANOVA Model of Difference in Number of Reported Cued Events Between the Two Diaries 67

Table A3.2 ANOVA Model of Difference in Average Event Duration of Cued Events Between the Two Diaries 68

Table A3.3 ANOVA Model of Difference in Average Event Description Length of Cued Events Between the Two Diaries 69

Table A3.4 ANOVA Model of Difference in Average Self-Rated Confidence of Cued Events Between the Two Diaries 70

Table A3.5 ANOVA Model of Difference in Response Time per Cued Event Between the Two Diaries 71

Table A3.6 ANOVA Model of Difference in Perceived Burden of Replying to Text Messages Between the Two Diaries 72

Table A3.7 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on the Number of Cued Events 73
Table A3.8 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Event Duration 74
Table A3.9 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Event Description Length 75
Table A3.10 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Self-Rated Confidence Score of the Reports 76
Table 4.1 Additional Measures of Data Quality for the Time Diary as a Whole 84
Table A4.1 ANOVA Model of Difference in Number of Reported Uncued Events Between the Two Diaries 100
Table A4.2 ANOVA Model of Difference in Average Event Duration of Uncued Events Between the Two Diaries 101
Table A4.3 ANOVA Model of Difference in Average Event Description Length of Uncued Events Between the Two Diaries 102
Table A4.4 ANOVA Model of Difference in Average Self-Rated Confidence of Uncued Events Between the Two Diaries 103
Table A4.5 ANOVA Model of How Cues Helped Respondents Remember Other Events 104
Table A4.6 ANOVA Model of Difference in Number of Reported Total Events Between the Two Diaries 105
Table A4.7 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on the Number of Uncued Events 106

List of Figures

Figure 2.1 The Proposed Memory Bookmarking Approach 10

Figure 2.2 Elapsed Time from Time of Signal to Time of Reply (in Minutes) 17

Figure 2.3 Simplified Diagram of the Data Collection System Used in the Main Study 20

Figure 2.4 Data Collection Process 23

Figure 2.5 Baseline Time Diary Flow 23

Figure 2.6 Second Time Diary Flow (Excluding Control Group) 25

Figure 2.7 Text Signals Sent by System and Hypothetical Response Examples 26

Figure A2.1 Time Diary Instrument of the Pilot Study (PC Version Only) 31

Figure A2.2 Mobile Interface of Baseline Time Diary (Main Study) 44

Figure A2.3 Mobile Interface of Second Time Diary (Main Study) 45

Figure 3.1 Data Quality Measures of Cued Events 53

Figure 3.2 Number of Cued "Watching TV and Playing Games" Reported by All Respondents 55

Figure 3.3 Moderating Effect of Cue Type on Data Quality Measures 57

Figure 3.4 Number of Cued Events Reported by Memory Capability 58

Figure 3.5 Average Event Duration of Cued Events by Memory Capability 59

Figure 3.6 Response Time per Cued Event (in Seconds) 60

Figure 3.7 Perceived Burden of Replying to Text Messages by Formality 61

Figure A3.1 At-home Events for Which Map Cues Are Not as Effective as Other Cues 77

Figure A3.2 Events for Which Map Cues Are Similar to Other Cues 78

Figure A3.3 Number of Cued Events by Education Level 80

Figure 4.1 Data Quality Measures of Uncued Events..... 86

Figure 4.2 Perceived Helpfulness of Cues by Cue Type 87

Figure 4.3 Number of Uncued Events Reported by Memory Capability 88

Figure 4.4 Perceived Helpfulness of Cues by Memory Capability 89

Figure 4.5 Response Time per Uncued Event (in Seconds) 90

Figure 4.6 Data Quality Measures of the Entire Time Diary (MB Conditions) 91

Figure 4.7 Data Quality Measures for the Entire Time Diary (Control Condition Only) 93

Figure A4.1 Data Quality Measures of Uncued Events by Cue Type 107

Figure A4.2 Perceived Work in Completing Entire Time Diary 108

Figure A4.3 Data Quality Measures for the Entire Time Diary 109

Abstract

Understanding human behavior is the goal of many social sciences, but accurately measuring the everyday behaviors of individuals is difficult. To collect information about and from individuals, researchers usually use retrospective survey questions; that is, questions about events that occurred in the past. Although these retrospective self-reports are prevalent, completing them places a considerable burden on respondents. According to a number of validation studies, recall may also be inaccurate. These recall errors can undoubtedly contribute to error in population estimates derived from survey data, particularly if respondents predominantly under- or over-report their activities, thereby leading to biased estimates.

In this dissertation, we propose and implement a new data collection methodology called memory bookmarking (MB). MB begins by collecting very minimal *in situ* information from respondents via texting. Texting for MB purposes does not require much input from respondents; it merely requires respondents to click on a web link to record their GPS location, take a single photo, or reply to a single text message at any signaled time. This *in situ* information is subsequently used as recall cues to prompt the respondents in a follow-up survey.

We conducted an MB-based pilot study using textual cues and a follow-up time diary to demonstrate the feasibility of this two-step data collection framework (Chapter 2) and help us design the experiment and system for the main study. In the main study, we tested the ability of two kinds of text cues, photo and map cues, to promote recall.

We found that such recall cues helped respondents to remember and report more cued events than they were able to during spontaneous recall without cues (Chapter 3). Moreover, the

events reported using cues were more fine-grained than the events reported without cues. Map cues require the least effort on the part of the respondents but provide the least amount of information for recall compared to the other types of cues (text cues and photo cues). Subjectively, respondents in this study were most confident about their recall when photo cues were present. This study provides evidence that the MB approach works more effectively for individuals whose memory capability is relatively low.

On the other hand, we found a neutral to negative effect on uncued events that were adjacent to cued events in the same time diary. Overall, we saw higher-quality reports of cued events (Chapter 3) but equivalent- or lower-quality reports of uncued events, making for a seemingly zero-sum game (Chapter 4). This make senses from the respondents' perspective; it seems that they were only willing to put a fixed amount of effort into the survey task. If they put more effort into some subtasks, then they put less effort into others. The resulting net impact on the time diary depended on the sum of the effects on cued and uncued events, which varied by individual. For instance, respondents with low memory capability benefited the most with respect to cued events but they were not affected when recalling uncued events, which resulted in a net gain in the total number of reported events. Other respondents showed a negative impact using this approach, which may be because of the particular way that the web instrument was designed and implemented. We provide detailed recommendations and future directions for implementing the MB approach for the time diary application.

Chapter 1 Introduction

Many essential quantitative characteristics of human society can be determined based on surveys of individual respondents. Social scientists measure these important features via survey questions, many of which are about events or activities that occurred in the past. When adopting retrospective questions, survey researchers have to depend on the autobiographical memories of respondents, which are prone to error (Bradburn et al., 1987; Tourangeau et al., 2000). Nevertheless, retrospective questions are still widely used in national surveys and form the basis of our knowledge about society. Why are retrospective questions so difficult and error-prone? By nature, humans are not good at memorizing and recalling things in the way that retrospective questions are generally asked. Humans do not organize their memories like computers do. Respondents cannot go to a folder or file labeled "yesterday" and read the memories saved there. When individuals recall memories, they are very likely to make errors. For instance, they may not encode every event of interest, they may not reconstruct events completely, or they may not be able to distinguish between similar events easily.

A substantial number of actions have been taken to address recall error and improve data quality. For instance, the life history event (LHE)/event history calendar (EHC) approach (Caspi et al., 1996; Belli, 1998) guides respondents to recall life events through a hierarchical structure; other aided recall approaches use contextual retrieval cues, temporal boundaries, and personal landmarks (Wagenaar, 1986; Barsalou, 1988; Menon, 1997; Robinson, 1986; Means and Loftus, 1991). There are also real-time data collection methods such as the ecological momentary assessment (EMA; Stone and Shiffman, 1994) and the experience sampling method (ESM; Larson

and Csikszentmihalyi, 1983). In addition, passive data collection methods from sensors on mobile or wearable devices can replace self-reports and thus eliminate the need for respondents' input (e.g., Eagle and Pentland, 2006; Doherty et al., 2017). This dissertation is one of many possible extended studies based on existing efforts to address the memory error problem in survey data collection. The newly proposed memory bookmarking (MB) approach requires the collection of minimal *in situ* information during a reference period. This information is then used to provide recall cues to help respondents reconstruct memories around the recorded time.

Memory Error Sources in Self-reports and Existing Methods to Reduce Memory Error

Research across the social sciences involves querying participants about everyday events. These events or episodes are part of what Tulving (1972) categorized as *episodic* memory, which "receives and stores information about temporally dated episodes or events, and temporal-spatial relations among these events" (p. 395). When asking such retrospective questions, researchers often have no choice but leave up to the participants how they search for relevant information about the target events, hoping they will locate and find the correct episodes or events and form an accurate answer based on what they recall. However, in reality, recall is not an easy task for anyone, especially when survey questions are not about the most memorable events in respondents' lives but rather about mundane and ordinary events. To address the recall problem, or why individuals can or cannot accurately recall episodic events, it is necessary to investigate how human memory works.

First, encoding is critical to recall. Encoding in the human memory system is the process of transforming the information a person receives from their senses to a memory unit that can be stored and retrieved. Survey questions frequently concern events that respondents did not encode

well or at all when they occurred (Lee et al., 1999). Even if the events have been encoded, the level or depth of processing at the initial encoding stage can affect how well they are remembered (Craik and Lockhart, 1972). The deeper the processing level, the more accurately the information is recalled. For instance, encoding the meaning of a word and relating it to similar words leads to better recall of that word (also known as *semantic processing*) than focusing on its phonological or visual properties, which leads to a shallower representation. Survey researchers who ask retrospective questions, particularly in cross-sectional studies, have no control over how respondents have encoded the target events. There are presently no survey techniques for improving recall that address the issue of poor encoding. This is because researchers usually have no idea about respondents' circumstances before they complete the survey interview or questionnaire. However, the proposed MB approach theoretically promotes deeper processing when memory bookmarks are collected. Chapter 2 explains the rationale behind this in more detail.

Second, elapsed time matters. The more time that has passed, the more individuals forget (see Tourangeau et al., 2000, for a review). Shortening the reference periods cannot fully solve the problem. If reference periods are shortened, it is necessary to add more waves of interviews in order to acquire the same amount of information about a person's behavior as before. The EMA and ESM methods represent extreme cases (Larson and Csikszentmihalyi, 1983; Stone and Shiffman, 1994), where participants only need to answer questions about the very moment that they are signaled (e.g., their activity, feelings, thoughts, and observations at and about the signaled moment). They are signaled several times at random for a week or longer so that the total sampled moments and their answers are able to ultimately capture all of the meaningful variations in their daily lives. The problems associated with this approach include: (1) the high burden of assessments repeated throughout the day if multiple questions are asked each time, as well as a lack of evidence

regarding when and how often the data collector should send signals to capture sufficient information about a respondent's daily life; and (2) the possible interference in a respondent's day.

Third, events are organized in a hierarchical structure. According to Conway's (1996) hierarchical structure theory, there are three levels of autobiographical memory: the lifetime period; general events; and specific events (also known as event-specific knowledge, similar to Tulving's episodic memory). The lifetime period is the top level of memory that records distinct periods of one's life. This is followed by general events and event-specific knowledge. General events, which comprise mid-level memory, are sets of associated events or a series of memories linked together by a theme (Barsalou, 1988). These are also called extended events or event sequences. The temporal and thematic relationship between adjacent events gives a person more chances to remember the events of a sequence, which can be recalled if any part of the sequence is accessible. The LHE/EHC technique uses the hierarchical structure of autobiographical memory. Both techniques encourage respondents to enumerate their important life periods and then follow a top-down order for retrieving more extended or episodic events (Caspi et al., 1996; Belli, 1998).

Fourth, events interfere with each other. Episodic events, in Tulving's definition, are stored in a structured hierarchy (Kolodner, 1985; Barsalou, 1988; Conway, 1996) and are typically grouped with other similar events. Memorable events are those events that are distinct from the events of the same kind, such as events that are unique and rare or important to one's life and have some emotional impact (see Tourangeau et al., 2000, for a review). To recall a single event other than the memorable ones, the respondent may need to distinguish this event from all other similar events by generating or searching the right indices (e.g., Kolodner, 1985). The indices are usually the distinctive properties of the events. Assuming no two events are the same, the indices can be any combination of features that distinguish two events from one another. If the indices are

inaccessible, the respondent will be unable to find the right path to that event in their long-term memory. Studies have explored using decompositional prompts to help individuals find the right indices (i.e., enumerating all possible events of a kind when answering survey questions) (Menon, 1997). This technique is currently used by the Consumer Expenditure Survey (CE; Bureau of Labor Statistics, 2015). Instead of expecting respondents to rely on free (i.e., uncued) recall of all recent purchases, CE provides a very detailed taxonomy of all possible purchases (indices) for individuals to recognize and follow. However, the researcher's taxonomy of the events or activities might not match how respondents classify things in their own memories. Furthermore, decompositional prompts substantially increase the survey length.

Fifth, recall failure may be a cue-dependent phenomenon (Tulving, 1974). Memory retrieval is the process of accessing and remembering information stored in long-term memory. An important argument that Tulving makes is that recall failure occurs when individuals do not find the right cue. In practice, recall cues are used as one of many aided recall tools. The contextual information about an event, which includes *who*, *what*, *where*, and *when*, may trigger respondents' episodic memories of specific events. The most effective recall cues are signals with properties that match the initial encoding operation (Tulving and Thompson, 1973) or information about what went on (Wagenaar, 1986; Brewer, 1988) and where the action occurred (Barsalou, 1988; Brewer, 1988). Wagenaar (1986) used contextual information recorded in his diaries to stimulate episodic memory; however, this kind of contextual information is not available in standard surveys. The common method of implementing contextual recall cues is to use additional prompts regarding *who*, *what*, *where*, and *when,* rather than to give respondents the contextual information.

Respondent Burden

When considering the reduction in memory error for any given approach, it is also necessary to consider if respondent burden is substantially increased. Some researchers have used response time for completing a survey question or a recall task as a proxy of respondent burden (e.g., Yan and Tourangeau, 2008; Collins and Quillian, 1969; Bousfield et al., 1954). Other researchers have examined individuals' perceived burden of completing a survey task (Belli et al., 2001; Belli et al., 2007). There may be a trade-off between data quality and respondent burden. Accordingly, when considering an approach designed to reduce memory error, researchers need to be aware of the impact of the approach on both data quality and respondent burden.

References

Barsalou, L. W. (1988). The content and organization of autobiographical memories. In In Neisser, U., & Winograd, E. (Eds.). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*, pp. 193–243. Cambridge University Press, New York.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, 6(4):383–406.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public opinion quarterly*, 65(1), 45–-74.

Belli, R. F., Smith, L. M., Andreski, P. M., & Agrawal, S. (2007). Methodological comparisons between CATI event history calendar and standardized conventional questionnaire instruments. *Public Opinion Quarterly*, 71(4), 603–622.

Bousfield, W. A., Sedgewick, C. H. W., & Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *The American journal of psychology*, 111–118.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical ques-tions: The impact of memory and inference on surveys. *Science*, 236(4798):157–161.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In Neisser, U. and Winograd, E., In Neisser, U., & Winograd, E. (Eds.). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*, pp. 21-90. Cambridge University Press, New York.

Bureau of Labor Statistics (2015). Consumer Expenditure Survey Quarterly Interview Survey Information Booklet. Retrieved from http://www.bls.gov/cex/2015_information_booklet_ce_305.pdf.

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., Smeijers, J., & Silva, P. A. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research*, 6(2):101–114.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247.

Conway, M. A. (1996). Autobiographical knowledge and autobiographical memories. In Rubin, D. C., editor, *Remembering our past: Studies in autobiographical memory*, pp. 67–93. Cambridge University Press.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., ... & Preece, S. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one*, *12*(2), e0169649.

Eagle, N., & Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, *10*(4), 255-268.

Kolodner, J. (1985). Memory for experience. *Psychology of learning and motivation*, 19:1– 57.

Larson, R., & Csikszentmihalyi, M. (1983). "The experience sampling method". New Directions for Methodology of Social and Behavioral Science, 15, 41-56.

Lee, L., Brittingham, A., Tourangeau, R., Willis, G., Ching, P., Jobe, J., & Black, S. (1999). Are reporting errors due to encoding limitations or retrieval failure? Surveys of child vaccination as a case study. *Applied Cognitive Psychology*, 13(1):43–63.

Means, B. & Loftus, E. F. (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology*, 5(4):297–318.

Menon, G. (1997). Are the parts better than the whole? The effects of decompositional questions on judgments of frequent behaviors. *Journal of Marketing Research*, 34(3):335– 346.

Robinson, J. A. (1986). Temporal reference systems and autobiographical memory. In D.C. Rubin (Ed.), *Autobiographical memory*, pp. 159–188.Cambridge, England: Cambridge University Press.

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*.

Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In Stone, A. A., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., and Cain, V. S., editors, *The science of self-report: Implications for research and practice*, pp. 29–47. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. & Donaldson, W., editors, *Organization of Memory*, pp. 381–402. Academic Press, New York.

Tulving, E. (1974). Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist*, 62(1):74–82.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352–373.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive psychology*, 18(2):225–252.

Yan, T. & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1):51–68.

Chapter 2 Conceptual Framework and Preliminary Data

2.1 Data Collection Framework

We developed and evaluated a new data collection method called memory bookmarking (MB) to facilitate recall for survey respondents. Figure 2.1 demonstrates how the MB approach works. There are two parties in this illustration: the researcher and the respondent. The bar at the bottom of the figure represents the information stored in the respondent's long-term memory, which is a "black box" for researchers. There are two stages in the proposed MB approach. The first stage is the collection of memory bookmarks. In Stage I, respondents are pinged, or signaled, via text messages and asked to provide information about their circumstances at the present moment. This *in situ* information is re-presented to them in Stage II as retrieval cues. For the rest of the dissertation, the terms "ping" and "signal" are used interchangeably, as are the terms "memory bookmark" and "retrieval/memory cues."

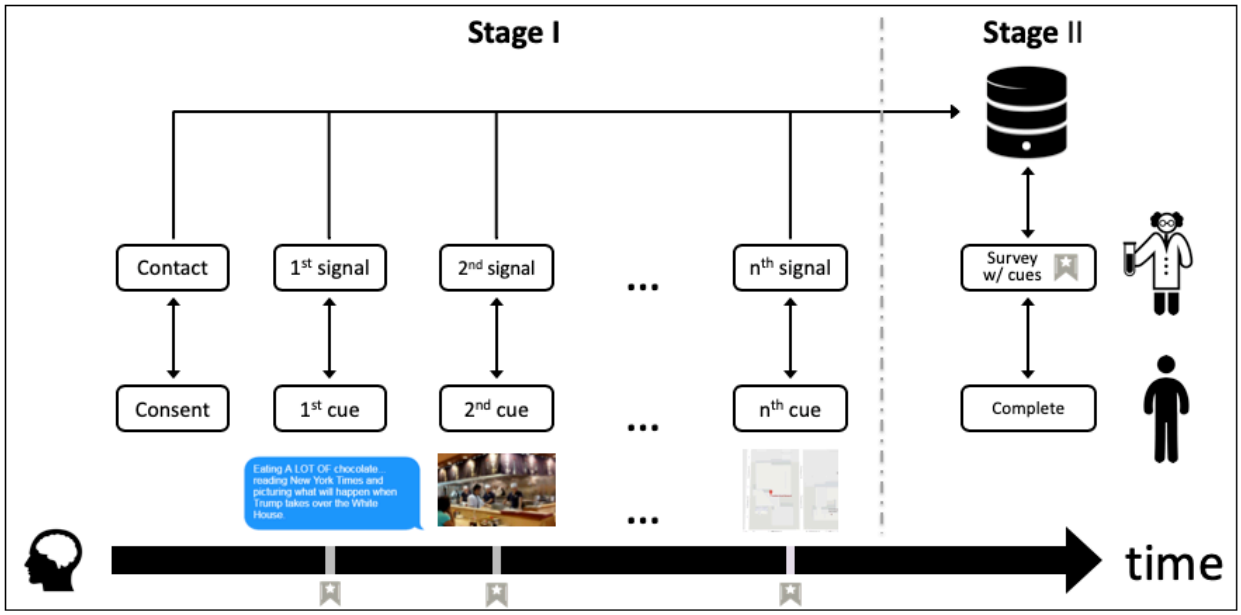


Figure 2.1 The Proposed Memory Bookmarking Approach
(The text, photo, and map shown in the figure are example cues.)

Each memory bookmark can be a textual self-report of the respondent's current circumstances, a photograph of the surrounding environment, a GPS location of the respondent at the signaled moment etc.. If textual, the respondent is instructed to provide either objective details or objective and affective aspects of their experience at that moment. The bookmark acts as an access point into the black box, depicted as gray slices in Figure 2.1. Stage II takes place on a subsequent day when the respondent is asked to recall aspects of the bookmarked events and other adjacent events occurred between bookmarked events. Presenting the collected memory bookmarks should not only enable the respondent to recall details of the cued event but should also—much like an actual bookmark allows a reader to see what is on the previous and next pages—promote recall of adjacent events. The memory bookmarks might also enable the respondent to recall other non-adjacent events from the time period.

Mechanism

Memory bookmarking approach helps individuals recall cued events

Memory bookmarks may affect the memory encoding process. A cue type that requires slightly more cognitive processing, such as generating a verbal description with details, likely deepens the level of processing (Craik and Lockhart, 1972), while a cue type that requires slightly less cognitive processing, such as sending one's GPS location using a smartphone, is less likely to affect the level of processing. During memory retrieval, the generated *in situ* information should closely match the information available at the time of encoding; this is known as encoding specificity (Tulving and Thompson, 1973). Regardless of cue type, the additional *in situ* information must include relevant contextual information about an event such as its physical surroundings, participants, and location. Therefore, the retrieval cue should have some impact on helping a person to recall that particular event (Wagenaar, 1986; Brewer, 1988).

Memory bookmarking approach helps individuals recall adjacent events

Given that all of the cues belong to a longer sequence of events on a particular day and the adjacent events should have temporal and causal relationships, we assume that these cues can provide anchors for recalling other adjacent events. For instance, if a respondent (Amy) recorded her activity on Saturday at 1 p.m. as "having a study group with Susan and Lee" and at 6 p.m. as "stuck on the highway," she might be able to use the two pieces of contextual information to remember the intervening activities. Amy must first quickly retrieve enough information to determine whether she traveled on the highway right after the study group. If the answer is no, she must think about how long the study group lasted, what she did immediately after the study group, and where she went. Given that at 6 p.m. she was on the highway and she was not driving directly from her meeting with Susan and Lee, she must start thinking about from where she left and what she was doing. This series of relevant retrieval tasks is automatically triggered based on the known information in a process similar to that of solving a crossword puzzle. The provided cues divide the entire full sequence into several shorter but bounded event sequences (e.g., Conway and Pleydell-Pearce, 2000; Thomsen, 2015). The cues themselves are also temporal boundaries with contextual information.

Previous studies have shown that listing relevant temporal boundaries or landmarks is a useful technique for a person attempting to retrieve memories of events that occurred near those boundaries (Robinson, 1986). Therefore, the MB approach might provide anchors for helping individuals recall events that occurred at adjacent but non-signaled times. This is similar to the event history calendar (EHC) approach, but the effectiveness of the cues might vary substantially. Unlike meaningful landmarks of a person's life in the EHC, the cues collected using the MB approach are random events captured during the day.

Applications

An intuitive application of this approach is a time diary, which allows researchers to learn about a person's activity and their experience of that activity on a particular day. Collecting memory bookmarks, such as verbal descriptions of signaled moments at several times during a particular day and feeding them back to the person, should allow the person to better remember the events that occurred at the signaled moments. The person may also remember the events that preceded and followed those events, eventually reconstructing the entire timeline of the day.

Consider another application. If a study is conducted to measure individuals' drinking behaviors over the course of a week, researchers could collect contextual information about when individuals are more likely to drink (e.g., nights and weekends). In Stage II, this information could be presented to respondents at the same time they are being asked questions like "How many drinks did you have in the past week?" and "How often did you go out for a drink last week?" By presenting the contextual information collected in Stage I (e.g., each of the respondents' locations each night during the past week), the respondents may find it easier to remember actual drinking or non-drinking occasions. In effect, the bookmarks convert a difficult free recall task into an easier cued recall task.

2.2 Focus of This Dissertation

In this dissertation, we evaluate three types of memory cues used in the MB approach—two types of textual descriptions (objective and affective), photo, and GPS location—that can be captured on smartphones and, in principle, easily implemented in mobile survey data collection. This dissertation primarily addresses the following five research questions:

1. Is the proposed MB approach feasible in a real-world mobile survey data collection setting? More specifically, do individuals consent to participate and complete the tasks as instructed? What is the completion rate of this form of data collection?
2. Does the proposed MB approach improve recall and thus data quality for the target events directly related to the cues?
3. Does the proposed MB approach help respondents to reconstruct a full sequence of events in a time diary?
4. Does the proposed MB approach introduce a substantial response burden?
5. Does the type of retrieval cue (textual description, photo, or map cue) make a difference in questions 2–4?

2.3 Pilot Study

Study Design

We conducted a small pilot study that used the MB approach to test its feasibility in a time diary setting. To be exact, we asked respondents to reply to all signals as instructed in Stage I and then, after an interval of 24 or 48 hours, respond to a time use survey in which memory bookmarks were embedded. We used this pilot study to inform our thinking about how to build a large-scale automatic data collection system and an effective mobile interface to collect such data.

In this pilot, we manipulated two experimental factors: doses of the cues (0 cues vs. 2 cues vs. 4 cues) and the length of the recall period (24 hours vs. 48 hours).

Table 2.1 Experimental Conditions of the Pilot Study

Factor II: Cue dosage	Factor I: Length of the recall period	
	24 hours	48 hours
Control - 0 cues	24 hours – 0 cues	48 hours – 0 cues
MB - 2 cues	24 hours – 2 cues *(at 10AM and 3PM)*	48 hours – 2 cues/day *(at 10AM and 3PM)*
MB - 4 cues	24 hours – 4 cues *(at 11AM, 2PM, 5PM, and 8PM)*	48 hours – 4 cues/day *(at 11AM, 2PM, 5PM, and 8PM)*

Given the limited capability of the survey platform we used for the pilot (Qualtrics.com), we tested the MB approach using only textual descriptions as cues. In Stage I, the respondents provided cues when signaled a few times over one or two days: "*What are you doing now? Please describe your circumstance in one text at the time you reply, including what you are doing, where you are, and whom you are with. [U-M time use study].*" In Stage II, each of the respondents received a time diary that embedded their text replies at the top of the time diary (see Appendix 2.1). The respondents in the 24-hours condition were asked to report their activities on the previous day, while the respondents in the 48-hours condition were asked to report their activities on the

previous two days. The time diaries were programmed using Qualtrics and required respondents to answer the following questions for each event (see Appendix 2.1):

(1) what activity they were doing (open-ended text box);

(2) when they started and when they finished doing that activity (open-ended text box);

(3) where they were (open-ended text box);

(4) who they were with (open-ended text box); and

(5) how happy were they on a scale from 0 to 6, with 0 being not happy at all and 6 being very happy (drop-down menu).

Feasibility

We recruited participants via Amazon Mechanical Turk, Craigslist, and an email invitation to staff members at the University of Michigan in October 2016. A total of 347 participants completed the screening survey. Among the 347 who signed up, 173 provided a cell phone number and confirmed their participation via a text message. Ultimately, 137 participants completed the entire task. The overall completion rate across the two recall periods was 79% (see Table 2.2).

Table 2.2 Completion Rate by Experimental Condition

		0 pings		2 pings		4 pings		Total	
		n	%	n	%	n	%	n	%
24 hours	Assigned	28		29		29		86	
	Completed *	22	79%	25	86%	27	93%	74	86%
48 hours	Assigned	29		29		29		87	
	Completed	21	72%	21	72%	21	72%	63	72%
Total	Assigned	57		58		58		173	
	Completed	43	75%	46	79%	48	83%	137	79%

* Respondents who replied to fewer than half of the signals were not allowed to continue onto Stage II (the time diary). Only a few respondents did not proceed to Stage II. The final completion rate was calculated by the number of people who completed the study divided by the number of people who confirmed their participation.

The completion rate varied slightly by cue dosage and recall period. The difference between the control, two-cue, and four-cue groups was minimal. As expected, the 24-hours group

had an overall higher completion rate (86%) than the 48-hours group (72%), which we attribute to survey fatigue. However, the four-cue group showed a similar or even slightly higher completion rate (83%) than the two-cue group (79%), indicating that the effort of replying to additional pings is negligible or at least does not reduce compliance.

Analyzing the timestamps of the text messages exchanged (signals from Qualtrics and event descriptions from the respondents), we found that most respondents were able to send an event description almost immediately after being signaled. Figure 2.2 shows the distribution of the lag between receiving and replying to a signal. The vertical dashed line marks the 99th percentile (120 minutes). The median response time for a signal was seven minutes.

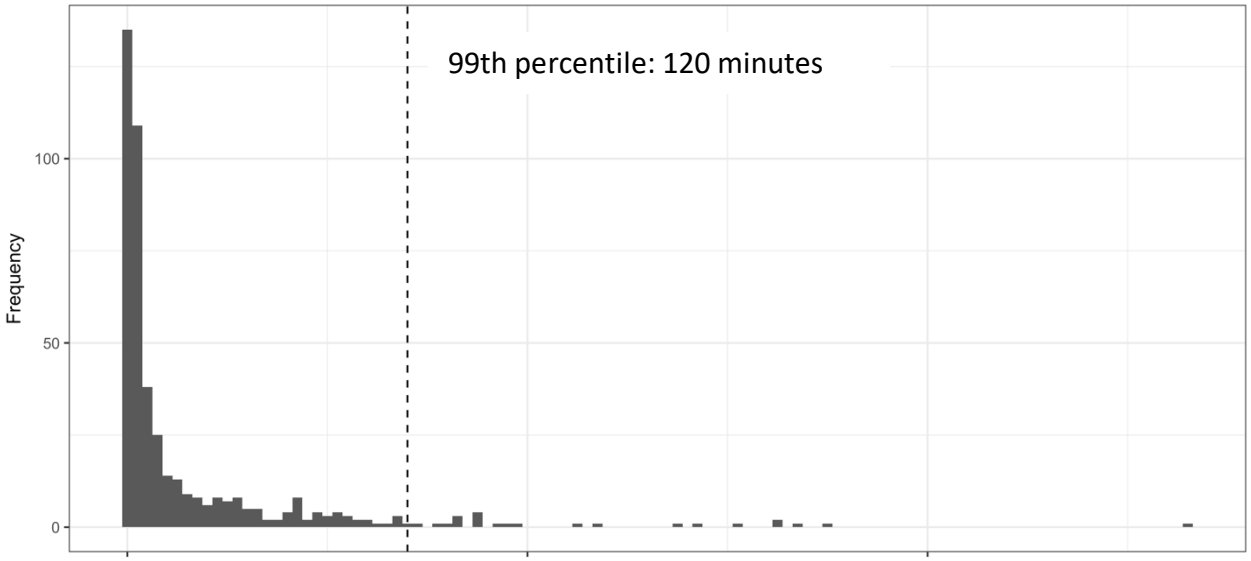


Figure 2.2 Elapsed Time from Time of Signal to Time of Reply (in Minutes)

Preliminary Findings and Implications for the Main Study

1. *The effect size is small; we should focus primarily on the high-dose condition.* In the pilot study, we mainly compared the average number of events reported in the time diary by respondents in each of the experimental groups. We hypothesized that the more events reported, the more effective the cues. Given the small sample size, the group means were quite similar and none of the differences were statistically significant. However, we did see some directional

evidence that the four-cue condition led to more recalled events. This is consistent with our expectation that the more memory cues that are provided, the larger the recall benefit. Moreover, respondents in the pilot study in the four-cue condition did not report that the task required substantially more effort than those respondents in the two-cue condition. Given a limited budget and a bigger effect size, we should *restrict comparisons to no cues vs. four cues*.

2. *We should reduce noise as much as possible.* In the pilot study, we were not able to balance respondents based on their demographics across all conditions. These external factors inevitably added more variation/noise to the group comparison. To reduce the impact of these external factors, we need to balance respondents across all conditions, especially on characteristics that are related to memory capability (e.g., age, education level). Another thought is to have a within-respondent experimental design, which would give us the most statistical power. Since the within-person variation is likely to be smaller than the variation between individuals, any large differences should be primarily due to the experimental manipulation.

3. *The day of the week should be considered as a blocking factor in the main study*. Although the difference between groups was small in the pilot study, we observed that responses were slightly different on weekdays and weekends. Given the irregularity of many individuals' weekend schedules, respondents may have been less able to report on a *typical* weekend day, requiring them to recall the events that actually occurred on those specific days.

4. *We should collect 24-hour data only if budget is an issue.* In theory, the more time that has elapsed, the more people are likely to forget, leading to greater improvement based on the MB approach. In the pilot study, we did not see a substantial difference between the 24-hours condition and 48-hours condition. The 48-hours condition also cost more to implement because

the completion rate was lower than that of the 24-hours condition (72% vs. 86%). We should collect only 24-hour data in the main study if the budget is an issue.

5. *Validation data would be valuable.* Thus far, we do not have a benchmark for validating self-reported answers. If possible, we should add an objective benchmark such as a GPS benchmark from a respondent's mobile phone to compare to the self-reported location collected as part of the survey data.

6. *Automated data collection is necessary*. It is impossible to manually launch such a MB-based study at a large scale with more types of memory cues. We need to build a data collection system that can effectively interact with the respondents to collect memory cues and later insert the cues into their follow-up time diaries.

7. *A mobile-friendly web survey is necessary*. In this pilot study, we explicitly encouraged respondents to complete the time diary on a desktop computer because of the limited formats available in Qualtrics. However, almost a third of the respondents completed the time diary on a smartphone even though it seems much more burdensome to scroll, read, and type on a smartphone to complete this task than it does on a computer.

8. *Cued events and adjacent events should be separated*. We need to separate cued events from adjacent events in order to answer Research Questions 2 and 3, as stated in the *Memory Bookmarking Framework* section. We need to determine whether the proposed MB approach facilitates recall of cued events and whether the proposed MB approach helps a respondent to reconstruct an entire day's sequence.

9. Other operational implications should be considered. (1) We should send respondents prompt messages and multiple reminders to motivate the respondents to finish. (2) We should reschedule as needed to accommodate respondents' schedules, which would result in a

higher completion rate. (3) We should set up a rule as to when and how the pings are randomly distributed throughout the day.

2.4 Main Study Design

Building on the implications of the pilot study, we developed a customized data collection system to implement the MB approach and a new web interface for collecting time diary data. This system has the ability to collect more than text cues provided via a short messaging service (SMS); it can also collect photographs via a multimedia messaging service (MMS) and GPS locations that can be converted into maps via a respondent's mobile web browser.

System development

The custom data collection system (Figure 2.3) is capable of: (1) requesting screening data from Qualtrics, which was the platform used for hosting and distributing the screening survey; (2)

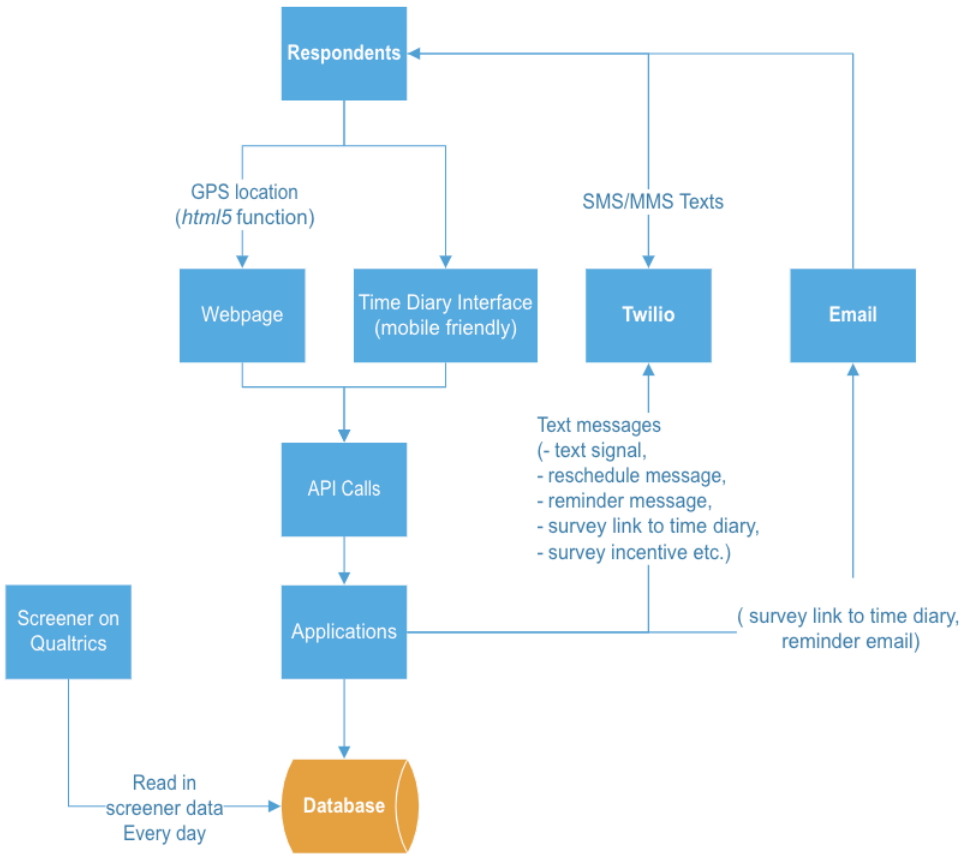


Figure 2.3 Simplified Diagram of the Data Collection System Used in the Main Study

collecting SMS text messages and MMS text messages from respondents via the third-party platform Twilio; (3) requesting and collecting GPS location data from respondents' web browsers using the *html5* function (a geolocation application programming interface [API] that is used to obtain the geographical position of a web browser from the user's device); (4) collecting time diary data with a new web interface that is mobile-friendly; and (5) interacting with participants for rescheduling and cancellation. In the figure, "Applications" represents the core of the system that manages experimental assignments and the rules for execution, data retrieval, and data storage.

Experimental Conditions

In the main study, we evaluated the MB approach using a randomized experimental design with five conditions: *Control* (no cues); *Text cues* (activity, location, and participants); *Text cues* (activity plus mood description); *Photo cues*; and *Map cues*.

Respondents were randomly assigned to one of the five conditions, so that the proportions were roughly the same across the conditions for six blocking factors: gender, age, education level, self-assessed memory (see the screening questionnaire in Appendix 2.3), the number of events reported in the baseline time diary (<12 events vs. 12+ events), and the day of the week (weekday or weekend day) on which events were queried. This was done to reduce the unexplained variance related to these six factors prior to the experiment and to make sure that the difference outcomes for the different conditions could only be attributed to experimental factors.

As part of the experiment, each respondent was measured twice. All respondents first completed the baseline time diary—a reconstruction of their events over the previous 24 hours—with no cues before they were assigned to an experimental condition. A week later, the respondents were signaled four times on a single day and asked to describe the events in which they were engaged. The respondents were to use a text, a photo, or geocoordinates, depending on the

condition to which they had been assigned. On the next day, they completed the second time diary in which the event descriptions they provided on the previous day were embedded as recall cues. The effect of cues versus no cues was generally assessed by measuring the within-respondent difference between the self-reports in the baseline and second time diaries. The effect of each cue type was also assessed by measuring the within-respondent difference in memory performance between the two diaries.

Overall Data Collection Process

Figure 2.4 illustrates in detail the data collection process for this experiment. We recruited volunteers who were 18+ years old and smartphone owners living in the United States using advertisements (see Appendix 2.2) that we posted on Amazon Mechanical Turk[1], Craigslist, Volunteer Science, and the Social Psychology Network. We also sent email invitations to a random sample of the staff at two large research universities in the United States: the University of Michigan and the University of Oklahoma. Volunteers completed a consent form (see Appendix 2.3) and provided their demographic information including phone number and email address as part of a screening questionnaire administered with Qualtrics.

The system read in screening data from Qualtrics every day and requested that each new participant on a particular day confirm their ability to participate. More specifically, each of the respondents needed to complete three tasks[2]: (1) send a text describing their current activity, (2) send a photo of their surroundings, and (3) click on a link to register their current geolocation[3]. If

[1] There was a $0.20 reward for completing a HIT task, which was a six-question survey that ultimately linked the respondent to the actual recruitment screener.
[2] **Text**: "Hi! Recently you agreed to participate in a time use study conducted at the University of Michigan. Please reply 'y' to confirm your participation." **Photo**: "In this study, you may sometimes be asked to send a photo. To be sure you can do this, please try sending a photo of your current surroundings. If your photo includes people's faces, please be sure to get their approval." **Map**: "In this study, you may sometimes be asked to report your current location. To be sure you can do this, please try clicking on the following link. {GEO-URL}"
[3] The accuracy of the GPS had to be within 0.5 miles.

they failed to send any of the three items in the active window (up to five hours)[4], they were screened out; that is, treated as ineligible and excluded from further data collection. Participants whose eligibility was confirmed were scheduled to complete the baseline time diary on either the next weekday or weekend day (approximately a 50-50 split[5]). A week later, the participants were contacted again to provide event descriptions (text, photo, GPS coordinates) for the second time diary. Once they had completed both, the respondents each received a $10 Amazon gift card.

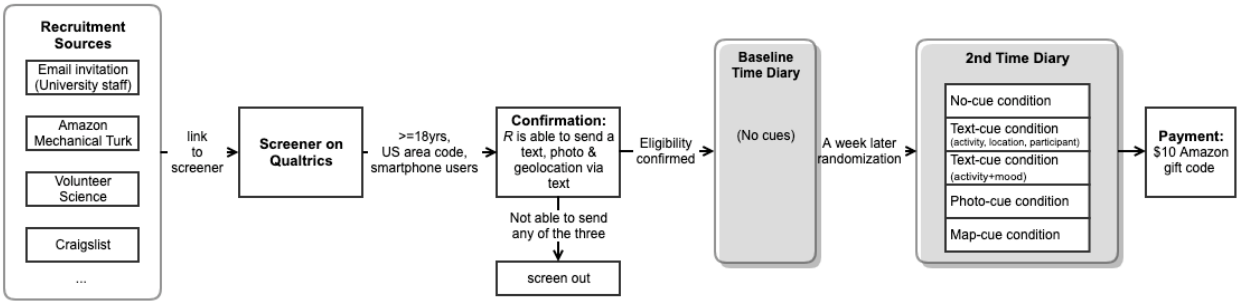


Figure 2.4 Data Collection Process

Baseline Time Diary

The baseline time diary was an online form spanning a single day and that was designed for completion on either a mobile device or desktop computer (see Appendix 2.4 and 2.5 for time diary instrument and its mobile interface). The system contacted all the participants, both those in the control group and those in the MB conditions, one day before the diary was distributed (Day

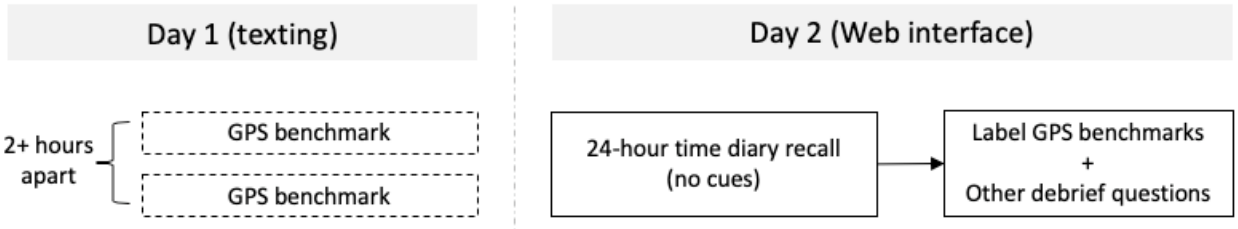


Figure 2.5 Baseline Time Diary Flow

[4] If they missed the active time window, they were prompted to reschedule. If they failed to reply to the rescheduling message within three days, their case was terminated.

[5] At the beginning of the data collection process, we assumed that the response rate would be lower on weekend days, so we allocated 40% of the respondents to weekdays and 60% to weekend days. As the data collection process continued, we found the response rate was almost the same. We thus reversed the two percentages later in the data collection.

1; see Figure 2.5) and asked if they were ready to begin[6]. If an affirmative reply was received by 11 a.m., the system proceeded, texting[7] each of the participants twice to record their geolocation benchmarks. The two text messages were sent before 11 p.m. that evening and were staggered so that they were sent at least two hours apart. These benchmarks were used to assess accuracy (see this chapter's *Preliminary Findings and Implications for the Main Study* section and Chapter 4). On the next day (Day 2), respondents received a survey invitation via text[8] and an email providing them with a link to the baseline time diary. This time diary asked them to report every event that had taken place from 4 a.m. on the previous day (Day 1) to 4 a.m. on the present day (Day 2). For each event a respondent reported, they were required to provide the following details: start time of the event, end time of the event, activity type (drop-down), open-ended description, where they were (drop-down), whether this event is part of their routine, perceived confidence score of their answer (0–10), how happy they were (0–10), how anxious they were (0–10), and how tired they were (0–10). Immediately after completing the time diary, each of the respondents was shown a debriefing section that included the two GPS benchmarks collected from the respondent on Day 1 using Google Maps. The respondent was asked to label these two locations using a drop-down menu. At the end of the debriefing section, we asked each respondent for their feedback about this study, including the burden of participating (see Appendix 2.4 for the questionnaire).

After the respondents completed the baseline time diary, the system randomly assigned them one of the five experimental conditions given the demographic strata into which they fell. The respondents were contacted again a week later to complete the second time diary. For instance,

[6] "Welcome to our time use study. Today you will receive a few text messages at different times during the day. Please reply to all messages when safe to do so. Please reply 'c' to continue, 'r' to reschedule, or 'stop' to opt out of the study."
[7] "Where are you now? Please click on the following link to report your current location. {GEO-URL}".
[8] "Hi! Please follow this link to complete the first survey of the Time Use Study: {J1-URL} We would appreciate it if you can complete the survey by the end of the day, today. We sent the same link to your email address, just in case you prefer to take the survey on your desktop/laptop computer."

if they had completed the baseline time diary on a Monday, they were scheduled to complete the second time diary on the next Monday. However, if they had missed the scheduled Monday, they were rescheduled for the next available weekday. The goal was to make sure that the respondents reported their activities on the same type of day (weekday or weekend) for both time diaries.

Second Time Diary

Participants in the control group completed a second time diary that was as same as the baseline time diary (i.e., no cues). The participants in the four MB groups (i.e., cues) went through a very different process (see Figure 2.6): not only did the system contact these respondents to collect two GPS benchmarks on Day 1, but it also sent another four text messages[9] throughout the day separated by at least two hours each. These text messages asked the respondents to report their current circumstances using the assigned cue type (i.e., text message, photo, or geolocation). Figure 2.7 shows the text messages that the system sent to the respondents in the four experimental conditions. The exact wording of the text messages for the four groups is as follows:

- *Text cue (activity, location, and participants)*

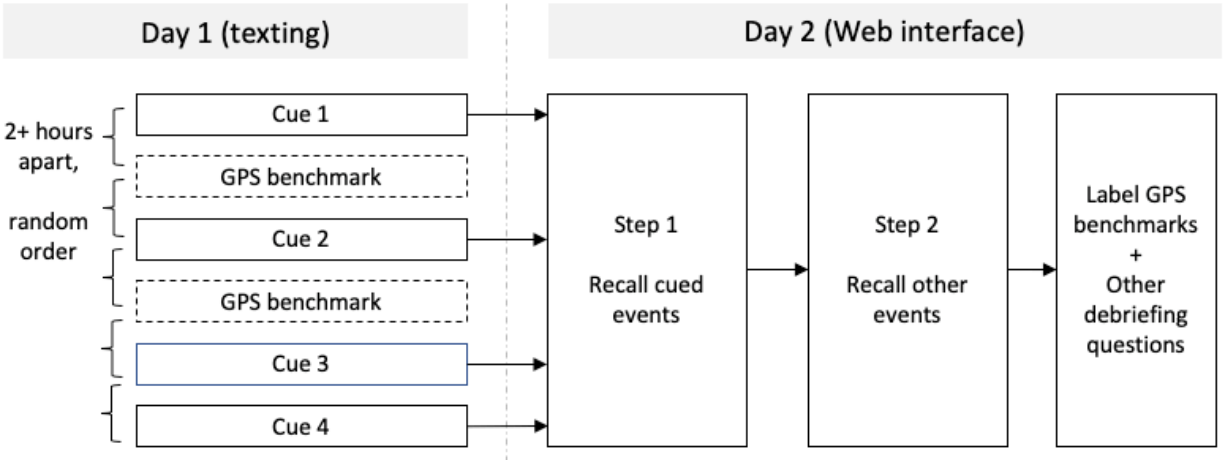


Figure 2.6 Second Time Diary Flow (Excluding Control Group)

[9] This was not completely random. From the time when a respondent agreed to start, the system sent a text signal every two hours. It was a random decision whether a signal was one of the four cues or one of the two GPS benchmarks.

Hi. Please tell us what you are doing now, including where you are, and who you are with. Please describe your situation at the time you reply not when you received the message if they are different.

- *Text cue (activity and mood)*

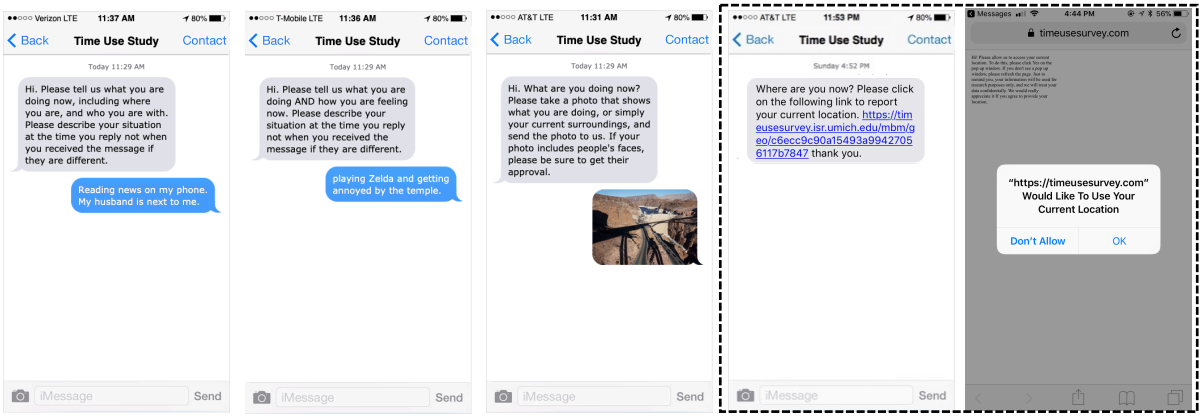
Hi. Please tell us what you are doing AND how you are feeling now. Please describe your situation at the time you reply not when you received the message if they are different.

- *Photo cue*

Hi. What are you doing now? Please take a photo that shows what you are doing, or simply your current surroundings, and send the photo to us. If your photo includes people's faces, please be sure to get their approval.

- *Map cue*

Where are you now? Please click on the following link to report your current location. {GEO-URL}



Text cue
(Activity, location, + participants)

Text cue
(Activity + feeling)

Photo cue

Map cue
(The respondent received a web link via text. Clicking on the web link prompted a dialogue asking for location access.)

Figure 2.7 Text Signals Sent by System and Hypothetical Response Examples

On Day 2, respondents in these four MB groups received a link to the second time diary. Like the baseline time diary, the second time diary was an online form spanning a single day and that was designed for completion on either a mobile device or desktop computer. Unlike the baseline time diary and the second time diary that respondents in the control group completed, the second time diary for respondents in the four MB groups included cues (text messages, photos, or maps constructed from the geocoordinates submitted by the respondents) preloaded at the appropriate times (see Appendix 2.5 for mobile interfaces). The second time diary for the four MB groups had two steps: respondents first needed to tap/click on each cue and fill out all information directly related to the cue, such as the start and end time of the event, activity type (drop-down), open-ended description, where they were (drop-down), whether this event is part of their routine, perceived confidence of their answer (0–10), and mood (how happy, worried, and tired from 0–10). Then they were instructed to report all the events that had happened between the cued events in order to complete the 24-hour time diary. After completing the time diary, the respondents completed the same debriefing questions as they did for the baseline diary, including labeling the GPS benchmarks and providing feedback about the task of completing the second time diary.

Participation

A total of 4,119 volunteers signed up for this study (completed the screening). Of these volunteers, 2,760 of them were eligible for participation in the study. Ultimately, 2,050 participants completed all the tasks and 1,976 respondents' data were used for analysis (we excluded data that were contaminated by issues such as unexpected system shutdowns and programming glitches). We compared the participation rate, or the number of people who completed both time diaries divided by the number of people who were eligible, across the five experimental conditions. The overall participation rate was 74.3%. The photo cue condition had a slightly lower participation

rate (71.5%) than the other cue conditions, which might be due to the slow rate at which MMS transmitted photos.

Table 2.3 Number of Completes per Condition of the Main Study

Experimental condition	# of assignment	# of complete	# for analysis	Participation rate (completes/assignment)
control	446	418	412	76.0%
Text cue (activity)	457	428	415	76.0%
Text cue (activity + feeling)	464	428	412	74.8%
Photo cue	447	394	377	71.5%
Map cue	425	382	360	72.9%

We examined the demographics of each group and found no significant differences among the experimental groups in terms of the blocking variables: gender, age, education, self-assessed memory, day of the week (weekday or weekend), and memory capability. The demographics are shown in Table 2.4.

Table 2.4 Demographics of Participants in the Five Experimental Conditions and Chi-square Tests for Differences Across Conditions (Main Study)

	n	%	X^2	d.f.	$Pr(>X^2)$
Age					
18–24 years old	370	19%			
25–34 years old	864	44%	6.067	12	0.9073
35–44 years old	481	24%			
45+ years old	261	13%			
Gender					
Male	612	31%	3.026	4	0.5535
Female	1364	69%			
Highest Education Degree					
Graduate degree	389	20%	2.413	8	0.9657
College degree	746	38%			
No college degree	841	43%			
Day of the Week					
Weekday	922	47%	0.323	4	0.9883
Weekend	1054	53%			
Self-rated Memory					
Excellent	537	27%	4.574	8	0.802
Very good	970	49%			
Good/Fair/Poor	469	24%			
Memory Capability					
12+ events in Baseline	1174	60%	1.713	4	0.7884
<12 events in Baseline	802	40%			

Note: Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

References

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In Neisser, U. and Winograd, E., In Neisser, U., & Winograd, E. (Eds.). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*, pp. 21-90. Cambridge University Press, New York.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological review*, *107*(2), 261.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684.

Robinson, J. A. (1986). Temporal reference systems and autobiographical memory. In D.C. Rubin (Ed.), *Autobiographical memory*, pp. 159–188.Cambridge, England: Cambridge University Press.

Thomsen, D. K. (2015). Autobiographical periods: A review and central components of a theory. *Review of General Psychology*, *19*(3), 294-310.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352–373.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive psychology*, 18(2):225–252.

Appendices

Appendix 2.1 Time Diary Instrument of the Pilot Study (PC Version Only)

Figure A2.1 Time Diary Instrument of the Pilot Study (PC Version Only)

Appendix 2.2 Main Study Advertisements

(Advertisement on Amazon Mechanical Turk and Craigslist, among other locations)
Short Screener of a Time Use Study ($10 upon completion of the actual study)

Researchers from University of Michigan are conducting a Time Use Study about how people use their time throughout the day in a digital era. Before the actual Time Use Study, we would like to ask 6 brief questions about your smartphone use. This should not take more than 2 minutes. If you are interested, you can find more details and sign up here:
https://umich.qualtrics.com/jfe/form/SV_5hCYnEqAITfQdJr.

You must be at least 18 years old and must live in the United States to participate. You must be the sole user of a smartphone. You must be able to send/receive text messages and take photos on your smartphone.

Note: If you are interested in participating the actual Time Use Study, please click on the link at the end of the task to find out more details. Upon completion of the actual Time Use Study, participants will receive a $10 Amazon gift card as a token of our appreciation.

(Email invitation)

Subject: Smartphone Users Needed for a Time Use Study ($10 Amazon Gift Code)
From: timeusesurvey@umich.edu

Greetings! We are conducting a Time Use Study at the University of Michigan and are seeking smartphone users to participate. We are investigating the ways in which mobile technology might help people tell researchers how they spend their time and how they feel during the day on surveys. Time use data make it possible for researchers to understand people's everyday behavior and decisions. We will provide a $10 Amazon gift code as a token of our appreciation for your participation.

If you are interested, you can find more details and sign up here:
https://umich.qualtrics.com/jfe/form/SV_5hCYnEqAITfQdJr.

We will randomly select a subset of eligible participants (*) from those who sign up. If you are selected, you will be contacted via text on your smartphone to confirm your ability to send text messages, photos, and your geo locations. In this study, you will need to respond to several text messages and to complete two time diaries online. If you have further questions about the study, please contact us at timeusesurvey@umich.edu.

This study has been approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board. If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researchers, please contact the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board, 2800 Plymouth Rd., Bldg. 520,

Room 1169, Ann Arbor, MI 48109-2800, (734) 936-0933 or toll free, (866) 936-0933, irbhsbs@umich.edu. Please refer to IRB number #HUM00112681.

**Note: Participants must be 18 years of age or older and the sole user of a smartphone; they must currently live in the US.*

Sincerely,
H.Yanna Yan, Ph.D. candidate and Frederick Conrad, Ph.D.
Institute for Social Research
University of Michigan, Ann Arbor

Appendix 2.3 Screener Questionnaire on Qualtrics (Main Study)

Time Use Study
#HUM00112681

Principal Investigator: H.Yanna Yan, Ph.D. candidate, Institute for Social Research, University of Michigan, Ann Arbor
Faculty Advisor: Frederick Conrad, Ph.D., Institute for Social Research, University of Michigan, Ann Arbor

You are invited to participate in a research study about how much time people spend sleeping, working, and doing other activities and how they feel at the time. The study is also designed to explore the feasibility of collecting time use data on mobile devices including using text messages.

If you agree to participate, you will be asked initially to complete this brief questionnaire. We expect it to take no more than three minutes.

We will randomly select a subset of eligible participants* from those who complete this recruitment questionnaire (this questionnaire takes about 3 minutes to complete). If you are selected, you will be contacted on your smartphone to confirm your ability to send text messages, media messages (a photo), and geo locations (via a web browser). Once you are confirmed, you will be contacted twice to complete two parts of the study; the second part will take place one week after you've completed the first part. Each part of the study lasts two days. On Day One, you will be contacted via text message to answer up to six questions. You need to answer these questions by sending a text message, a photo or a current location. We expect the time and effort taken to respond to these text messages is minimal – less than 1 minute per message. On Day Two, you will receive a link to a time diary to report your daily activities. We expect the average time for completing the time diary to be 20 minutes.

Once you have completed the study, you will receive a $10 Amazon gift code as a token of our appreciation for your participation. By "you have completed the study" we mean: a) you've answered at least half of the text messages on Day 1; b) you have finished the time diary by the end of Day 2; c) you have fulfilled a) and b) for both parts of the study.

You will incur normal texting and cellular data fees when you participate in the study. The exact cost will depend on your plan with your cellphone carriers.

By participating in this study, you will be helping social scientists better understand how people allocate their time to different activities and how people feel while doing different activities. Furthermore, you will also be helping researchers identify and explore new ways of collecting time use data on mobile devices including using text messages. Although you may not receive direct benefits from your participating, others may ultimately benefit from the knowledge obtained in this study.

We don't believe that participating in the study will create any risks for you or that you will experience any discomfort by participating. We will not share your answers with anyone outside

the project. There is never perfect security in mobile communication but the risk of someone intercepting your answers is extremely low and no higher in our survey than when you use your phone every day.

We will store all the data, no matter whether your participation is partial or complete, at Institute for Social Research, University of Michigan for future analysis and publications. However, You cell phone number and email address will be destroyed once the data collection is over, so we will not be able to link your survey responses to you.
We expect that the general results of this study will be published in scientific articles in scientific journals, but nothing about individual answers.

It is totally up to you to decide to be in this research study. Participating in this study is voluntary. Even if you decide to be part of the study now, you may change your mind and stop at any time. You do not have to answer any questions you do not want to answer. If you decide to withdraw before this study is completed, we will still store your partial answers. However your cellphone number and email address will be destroyed once the data collection period is over. Your participation may also be terminated by the PI if you: a) have not confirmed your participation via text messages; b) have not responded to half of questions on Day One for both parts of the study.

If you have questions about this research, you may contact H.Yanna Yan and Frederick Conrad at timeusesurvey@umich.edu.

This study has been approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board. If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researchers, please contact the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board, 2800 Plymouth Rd., Bldg. 520, Room 1169, Ann Arbor, MI 48109-2800, (734) 936-0933 or toll free, (866) 936-0933, irbhsbs@umich.edu. Please refer to IRB number #HUM00112681.

*Note: Participants must be 18 years of age or older, and the sole user of a smartphone. Participants must currently live in one of the following US Time Zones: Hawaiian Time Zone, Alaskan Time Zone, Pacific Time Zone, Mountain Time Zone, Central Time Zone and Eastern Time Zone. Participants must have a valid US cell phone number for texting.

consent. Do you agree to participate in the Time Use Study?

By clicking "Yes, I agree", you are agreeing to be in this study. Make sure you understand what the study is about before you click. We suggest you printing out a copy of the consent page for your record. If you have any questions about the study, you can contact the study team using the information provided above.

- 1-Yes, I agree.
- 2-No, I don't agree (END).

Timezone. Which time zone are you in? If you are traveling, please select the time zone that you will be in for the next two weeks.

- 1 – Hawaiian Time Zone
- 2 – Alaskan Time Zone
- 3 - Pacific Time Zone
- 4 - Mountain Time Zone (except Arizona)
- 5 – Mountain Time Zone (Arizona only)
- 6 - Central Time Zone
- 7 - Eastern Time Zone
- 8 - I am not or will not be in any of the listed time zones for the next two weeks (END, END message)

Gender. Do you identify yourself as ...

- 1 - Male
- 2 - Female

Age. What is your age?

- 1 -- Under 18 years (END, END message)
- 2 -- 18-24 years old
- 3 -- 25-34 years old
- 4 -- 35-44 years old
- 5 -- 45-54 years old
- 6 -- 55-64 years old
- 7 -- 65 years or older

Education. What is the HIGHEST degree or level of school you have COMPLETED? (Modified Q.AG6 NSFG 2011-2013)

- 1 -- Less than high school
- 2 -- High school graduate or GED
- 3 -- Some college without a degree
- 4 -- 2-year college degree (e.g., Associate's degree)
- 5-- 4-year college graduate (e.g., BA, BS)
- 6-- Graduate or professional school

Employment. In the last seven days, did you have a job either full or part time?
Please include any family farm or business, and any job from which you were temporarily absent.
(ATUS 2016 modified from FABS)

- 1 - Yes
- 2 - No
- 3 - Retired
- 4 - Disabled
- 5 - Unable to work
- 6 - Other

Memoryself. How would you rate your memory at the present time? (D101, HRS2014)

- 1 - Excellent

- 2 - Very Good
- 3- Good
- 4- Fair
- 5- Poor

Kids. Do you have any children under age 18 who are currently living with you?
Please do not count any children who are currently attending school, such as boarding school or college.

- 1-Yes
- 2-No

Hispanic. Are you of Hispanic, Latino, or Spanish origin? (CPS 2017 HSPNON)

- 1 - Yes
- 2 - No

Race. What is your race? You may choose one or more options. (CPS 2017 RACE)

- 1 - White
- 2 - Black or African American
- 3 - American Indian or Alaska Native
- 4 - Asian
- 5 - Native Hawaiian or Other Pacific Islander

Marital. Are you now married, widowed, divorced, separated or never married? (CPS 2017 MARITL)

- 1-Married
- 2-Widowed
- 3-Divorced
- 4-Separated
- 5-Never married

Cellnum. We will contact you via text and email for the Time Use Study. What is your US cell phone number?

_____

Email. What is your email address?

_____

Source. Where did you find this study? e.g., Amazon Mechanical Turk etc.

(END message)
Sorry, participants must be 18 years of age or older, and the sole user of a smartphone. Participants must currently live in one of the following US Time Zones: Hawaiian Time Zone, Alaskan Time Zone, Pacific Time Zone, Mountain Time Zone, Central Time Zone and Eastern Time Zone. Participants must have a valid US cell phone number for texting. Thank you for your interest!

(SUCCESS)
Thank you for completing the recruitment survey! We will select a random group of participants. If you've been selected, we will contact you via text message to confirm your enrollment. Thank you!

Appendix 2.4 Time Diary Instrument of the Main Study

Page 1 - Welcome page

(Baseline Time Diary)

Welcome!

Thank you for participating in the Time Use Study. This is your first survey. We will contact you again a week later about the second survey, and will send you a $10 Amazon Gift Card shortly after you complete the second survey.

Your responses are very important to us and will be used for research purposes only. We want to know how much time people spend sleeping, working, and doing other activities and how they feel at the time. We will treat your data confidentially: your phone number and email address will not be stored with or linked to your responses.

Again, we truly appreciate your participation.

(Second Time Diary)

Welcome!

Thank you for staying with us so far. This is your second survey. We truly appreciate all of your effort. We will send you a $10 Amazon Gift Card via text and email shortly after you complete this survey.

Your responses are very important to us and will be used for research purposes only. We want to know how much time people spend sleeping, working, and doing other activities and how they feel at the time. We will treat your data confidentially: your phone number and email address will not be stored with or linked to your responses.

Again, we truly appreciate your participation.

Page 2

We would like to know how you felt yesterday, [day, date]. Please rate each feeling on the scale from 0 to 10. A rating of 0 means that you did not feel that way at all; a rating of 10 means that you felt that way very strongly.

- Happy
- Worried/Anxious
- Tired

Page 3 Diary instruction

(Baseline Time Diary)

Please think of your day as a continuous series of episodes/activities like scenes in a film. The episodes/activities people identify usually last between 5 minutes and 2 hours. An indication that an episode has ended might be going to a different location, starting another activity, or interacting with other people.

Please do your best to report ALL activities that last more than 5 minutes, even if they take place during a longer activity. For instance, if you grab a coffee downstairs for just 5 minutes during work, please report "went to coffee shop" as well.

OK, I'll try my best. (This is a button)

(Second Time Diary)

Yesterday we collected several [text messages/images/locations] from you. We will show you these [text messages/images/locations] on the next page to help you remember all the activities that occurred yesterday. We call them "memory cues."

Just like last time, please think of your day as a continuous series of episodes/activities like scenes in a film. The episodes/activities people identify usually last between 5 minutes and 2 hours. An indication that an episode has ended might be going to a different location, starting another activity, or interacting with other people.

Please do your best to report ALL activities that last more than 5 minutes, even if they take place during a longer activity. For instance, if you grab a coffee downstairs for just 5 minutes during work, please report "went to coffee shop" as well.

OK, I'll try my best (This is a button).

Event Page

- What time did you start doing this activity?
  HH MM AM/PM (dropdown menu)

- What time did you end doing this activity?
  HH MM AM/PM (dropdown menu)

- Please briefly describe this activity (text box)

- What were you doing? (Please select)
  - Sleeping
  - Washing, dressing & grooming oneself
  - Eating and drinking
  - Traveling
  - Working or studying
  - Preparing food or cleaning up
  - Housework
  - Shopping including groceries
  - Watching TV/movies and playing games
  - Doing sports/exercising
  - Socializing w/ others, telephone calls
  - Caring for & helping others including kids
  - Other
  - Don't remember

- Where were you? (Please select)
  - My home or yard
  - My workplace or school
  - Restaurant/bar
  - Grocery store/other store/mall
  - Gym/health club
  - In transit (in car, bus, train, subway, boat, airplane etc.)
  - Hospital/Medical clinic/Doctor's office
  - Other place
  - Don't remember

- Is this activity part of your daily/weekly routine?
  - Yes
  - No

- On a scale from 0 to 10, how confident are you about your answers for this activity?

- We would like to know how you were feeling during this activity. Please rate each feeling on the scale from 0 to 10. A rating of 0 means that you did not feel that way at all; a rating of 10 means that you felt that way very strongly.
  - Happy
  - Worried/Anxious
  - Tired

Debriefing Page

You are almost done. We collected two locations from your smartphone on [day, date], and we need your help labeling these locations.

(2 GPS map shown on a map)

- Where is this? (Please select)
  - My home or yard
  - My workplace or school
  - Restaurant/bar
  - Grocery store/other store/mall
  - Gym/health club
  - In transit (in car, bus, train, subway, boat, airplane etc.)
  - Hospital/Medical clinic/Doctor's office
  - Other place
  - Don't remember

- Additional comments: (text box)

Debriefing Page 2

Using a scale from 0 to 10…

- Yesterday you received several text messages asking you to reply to some short questions. How annoying were the text messages? 0 = not annoying, 10 = very annoying.

- How much work was it to reply to those text messages? 0 = not much work, 10 = a lot of work.

- When you filled out your diary, how much did reading the memory cues help you remember OTHER activities? 0 = did not help much, 10 = helped a lot. (cue groups only)

- How much work was it to complete the time diary? 0 = not much work, 10 = a lot of work.

- How much fun was it to participate in this study? 0 = not much fun, 10 = a lot of fun.

- Was the way you completed the second time diary any different from the way you completed the first? (no-cue group, second wave only)[10]
  - Yes
  - No

- If Yes to the previous question, please briefly tell us how. [text box]

- Did the memory cues affect the way you completed the second time diary compared to the first? (cue-group only, second wave only)
  Yes
  No

- If Yes to the previous question, please briefly tell us how. [text box]
- Are you willing to be contacted another time for a follow-up study or a similar Time Use Study? (second wave only)
  - Yes
  - No

- Do you have any other feedback about this study? (text box)

Thank you Page

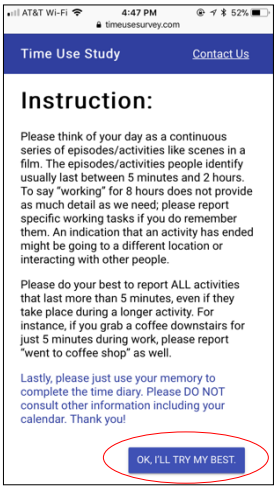
(first wave)
Thank you for submitting your survey. We will contact you again in about a week.
Now please close your browser window.

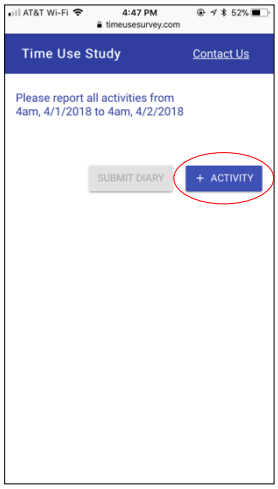
(second wave)
Thank you for participating in the Time Use Study. We will send you a $10 Amazon gift code via text and email shortly.
Now please close your browser window.

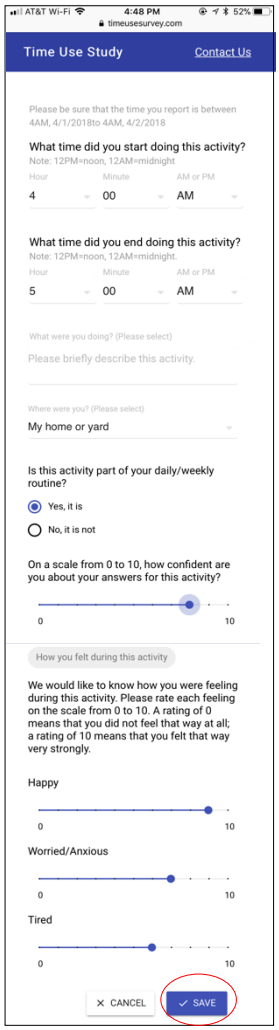
[10] Only asked during the second half of data collection.

Appendix 2.5 Mobile Interface of Time Diaries (Main Study)

Figure A2.2 Mobile Interface of Baseline Time Diary (Main Study)







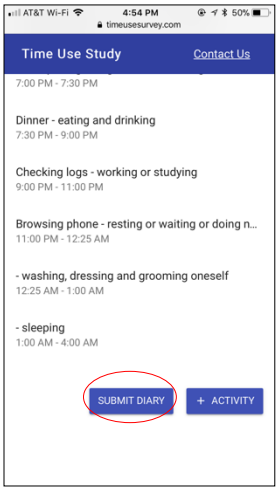
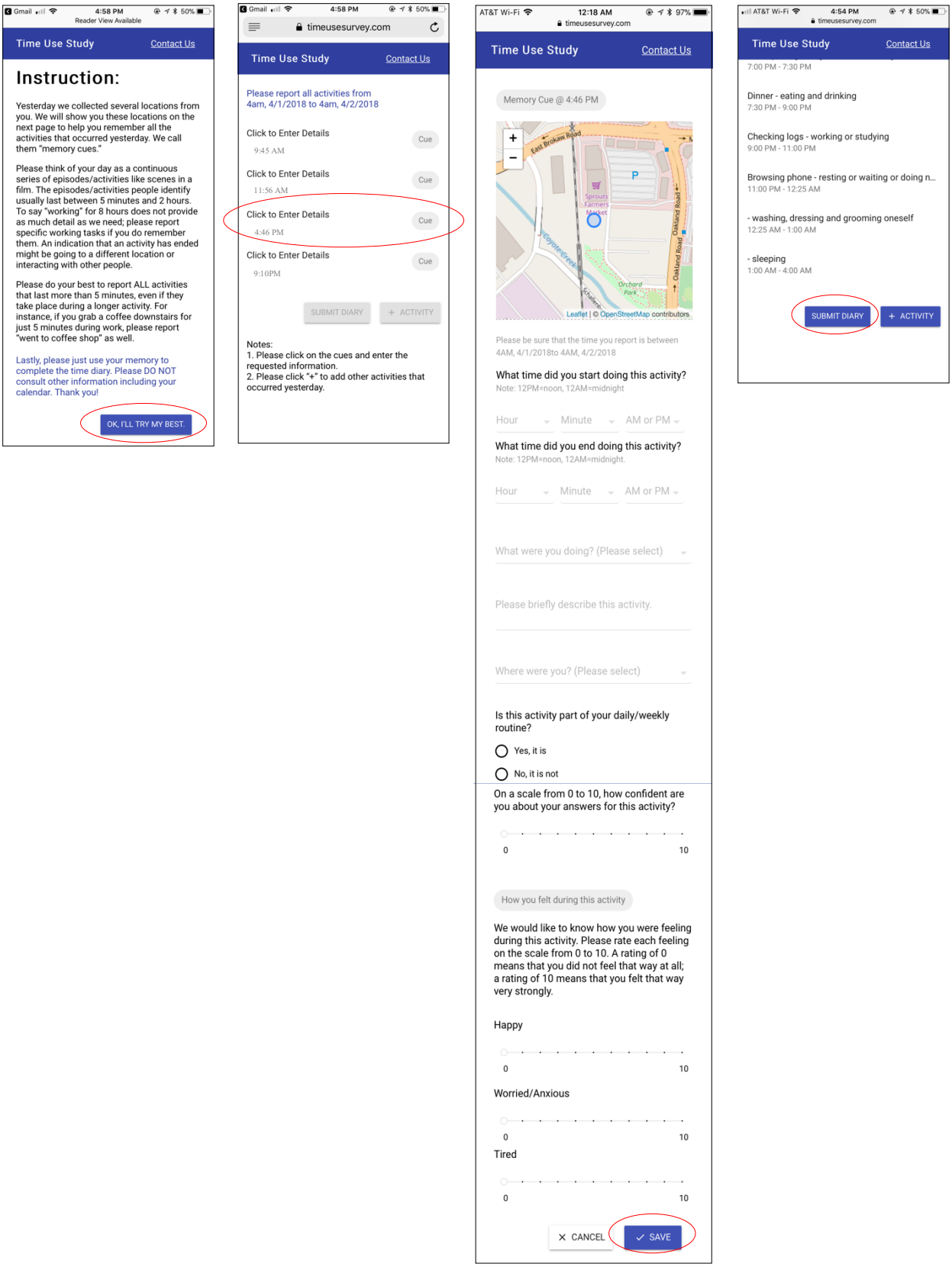


Figure A2.3 Mobile Interface of Second Time Diary (Main Study)



Chapter 3 Using Text, Photo, and Map Cues via Smartphone to Promote Recall

3.1 Introduction

Tulving (1974) proposed that recall failure may be a cue-dependent phenomenon. Recall failure, then, occurs when an individual does not find the right cue. The ideal recall cue is the information that most closely matches the information that was available at the time of encoding, which is known as encoding specificity (Tulving and Thompson, 1973). Although these findings were derived mainly from laboratory experiments involving recalling words, several studies have extended the use of recall cues to personal events in autobiographical memory (Wagenaar, 1986; Brewer, 1988). Despite their small sample sizes (either the researcher himself or a small group of students), these studies demonstrate that contextual information such as activity, location, and participants are effective cues for recalling one's personal events. Related to the encoding specificity phenomenon is mood dependent memory. According to the literature (e.g., Blaney, 1986), individuals recall target personal events more easily if they are in the same mood as they were the day that the events occurred (e.g., Eich et al., 1994).

Two pilot studies have tested the idea of collecting contextual information and re-presenting it later as retrieval cues (Hodges et al., 2006; Kelly et al., 2015), but no study has implemented and leveraged this idea in actual survey practice. For instance, one usability study (Hodges et al., 2006) tested a novel device called SenseCam—a sensor-augmented wearable stills camera—on a patient who had been diagnosed with limbic encephalitis and bilateral cell loss in the hippocampus that resulted in significant memory problems. When reviewing the photos taken by the SenseCam, the patient was able to recall more life events than they were when they did not

have photos. The effects on recall were substantial in both the short term (two days after the event) and long term (one month after the event). A feasibility study focusing on a time diary kept by healthy volunteers similarly examined the use of a wearable device to take photos of subjects' daily lives (Kelly et al., 2015); these photos were then presented as memory aids in prompted follow-up interviews. The researchers found that image-prompted interviews produced more self-reported events.

3.2 This Study

Given that external and internal contextual cues are effective in helping individuals recall personal events, how can researchers collect and use such cues in surveys that require the recall of daily events, such as time use surveys? This question cannot be fully answered by the aforementioned existing feasibility studies. Even less is known about how *different* kinds of recall cues might have an impact. This study helps to fill this research gap by collecting contextual information via individuals' smartphones and inserting this collected information into follow-up surveys as recall cues. To do so, we use the proposed memory bookmarking (MB) framework. The actual implementation is described in detail in Chapter 2. In this study, we test four different type of recall cues: textual descriptions of an event (the activity, location, and participants); textual descriptions of an event and the respondent's mood at the time of the event; a photo taken by the respondent during the event; and a map showing the location of the event based on geocoordinates recorded by the respondent's smartphone.

Hypothesis 1. Recall cues that are part of the memory bookmarking methodology promote better recall and reporting of personal activities compared to spontaneous (free) recall. Previous literature has provided adequate laboratory findings in this regard, but the focus of this study is to

implement the idea of contextual recall cues via smartphones and test the cues' usefulness in a web survey scenario.

Hypothesis 2. The effect of cues varies by cue type, and photo cues are the most effective cues. Both types of text cues remind participants of what they thought were salient features of the event at the time; they do not contain any information that was not explicitly entered by the participants. Map cues, although they are more objective than self-entered text, provide no information besides that of the location. Photo cues are as objective as the map cues. Even though a photo does not contain a map location, the physical environment shown in the photo should reasonably suggest or easily remind participants of the location. Beyond that, a photo may capture a larger context than a text cue or map with vivid details, probably more than the participant was aware of at that moment. In the previously mentioned feasibility study, Hodges et al. (2006) compared the focal patient's recall when prompted by written diaries and photos taken by SenseCam, finding that recall is much better if prompted by photos. However, the results of the study were more qualitative than quantitative.

No other empirical studies have directly investigated if and how different types of retrieval cues affect autobiographical recall. However, Brewer (1988) and Rubin et al. (2003) both found in their laboratory experiments that relived (also referred to as recollected/remembered) memories almost always involve strong visual images during the retrieval process. This gives us some evidence that visual memory aids are effective in stimulating the retrieval process. In addition, a number of studies have shown humans' ability to memorize and accurately recall an enormous quantity of pictures in detail (see Schurgin, 2018, for a review). For instance, one study showed that humans' picture recognition is extremely high compared to word and sentence stimuli, even after several days have passed since the humans' exposure to the pictures (Shepard, 1967). Another

study showed that people can memorize a large quantity of pictures with the ability to recognize the details contained in the pictures (Brady et al., 2008). These findings do not directly prove that visual cues are better but provide evidence of humans' exceptional ability to recognize the details of past visual stimuli from their own lives and to distinguish between them, thus suggesting the potential of photo cues.

Hypothesis 3. Cues provide greater benefit to individuals who have low memory capability. By memory capability, we mean the ability to store and retrieve events from autobiographical memory. The reasoning behind this is that there is more potential for these individuals to improve their memory compared to individuals who can remember past events very well. Those with strong recall abilities may perform near the ceiling without cues and so cannot demonstrate additional benefit from the introduction of cues.

Exploratory research question 1. Does additional affect information in text cues promote even better recall? Mood-dependent memory literature (e.g., Blaney, 1986; Eich et al., 1994) shows that individuals are more likely to remember target events when their mood while recalling the event matches their mood when the event occurred. Although we are unable to stimulate different moods in a survey setting, we can collect additional affect information as part of text cues and present it to the respondents. In this study, we explore whether the affect information in text cues is more or less effective in stimulating recall than other physical contextual information such as an event's location and participants.

3.3 Methods

There were four experimental (cue) conditions with approximately 400 respondents in each: *Text cue (activity, location and participants); Text cue (activity and mood description); Photo cue; and Map cue.* Randomization is balanced across the levels of six blocking factors: day

type (weekday or weekend day), gender, age, education level, self-assessed memory, and the number of events reported in the baseline time diary (<12 events vs. 12+ events). A detailed description and demonstration of the *Experimental Design* can be found in Chapter 2.

Measures of Data Quality and Response Burden

We compared data quality and respondent effort between two time diaries and across experimental groups using the indicators shown in Table 1: number of reported events, average event duration, description length, self-rated confidence score, perceived effort taken to respond to cues, and response time spent per event. Some of these indicators have been used in time diary assessments or other studies; the remaining ones were invented for this study. We acknowledge that none of the indicators is perfect, and they are unlikely to directly reflect the memory quality but do reflect the quality of the report as a whole. Researchers may be able to trace or monitor all aspects of a person's life in a real-world survey setting in the future; however, these are the best possible indicators available to us now.

Table 3.1 Measures of Data Quality and Effort

Measure	**Type**	**Indication**	**Statistical test**	**Previous literature**
Number or reported events	Continuous	More events means better memory and report.	1. **Main effect** of the cues: - T-test for testing difference (Δ) between baseline and second time diaries. 2. **Moderating effect** of the cue type and other factors: - Dependent variable is the difference (Δ). - Independent variable includes moderators and other blocking factors. - ANOVA F-test - Parameter estimate testing in linear regression.	Fisher and Gershuny, 2013
Event duration (in minutes)	Continuous	Shorter event duration means the events are more fine-grained.		New
Description length	Continuous	Longer description means better memory and report.		New
Self-rated confidence score	Integer from 0 to 10, treated as continuous	Higher score means more certainty.		Brewer et al., 1988; Brewer, 1996
Perceived effort taken to respond to cues	Integer from 0 to 10, treated as continuous	Higher rating means more effort.		New
Response time spent per event (in seconds)	Positive integer	Longer time means more effort.		Yan and Tourangeau, 2008

Analysis Approach

In the second time diary, respondents were given one kind of cue (one of the two types of textual description, photos, or maps) with timestamps to help them report the event that had occurred at that moment they had been signaled. We compared the cued events (i.e., events that occurred at the signaled times)[11] to their temporal counterparts captured in the baseline time diary (i.e., events whose reported time of occurrence spanned the time of each signal in the second time diary). Note that the counterparts shared only their time of occurrence with that of the actual cued

[11] In this paper, a signaled time is the time that the respondent replied to the text message signaling them to provide a cue.

events in the second time diary. We assumed that individuals' daily activities are similar from week to week (e.g., they run errands on Saturday or go bowling on Wednesday evening), making it reasonable to compare the two measures on the same day of the week to test the effect of the cue. For the sake of simplicity, in the following sections, we use "cued events" to indicate the cued events in the second time diary and their temporal counterparts in the baseline time diary.

When assessing the effect of cues on recalling the events underway when participants replied to their signals, we compared the quality and effort measures (see Table 1) of the cued events across the two time diaries. We examined: (1) if there was a significant difference between the two time diaries regardless of cue type and (2) if the difference (Δ) varied by cue type or other factors. Data management, statistical testing, modeling, and figures were completed using R.

3.4 Results

Overall Effect on Data Quality (H1)

Figure 3.1 shows the group means for the four quality measures for both time diaries. Regardless of the cue type, respondents reported 0.5 more cued events than temporally analogous events in the baseline study (t(1563) = 23.95, p-value =<0.001; see Figure 3.1a). Moreover, the average event duration was 37.3 minutes shorter (t(1563) = -11.36, p-value <0.001; see Figure 3.1b). Together, this suggests that cues promote recall of more fine-grained events.

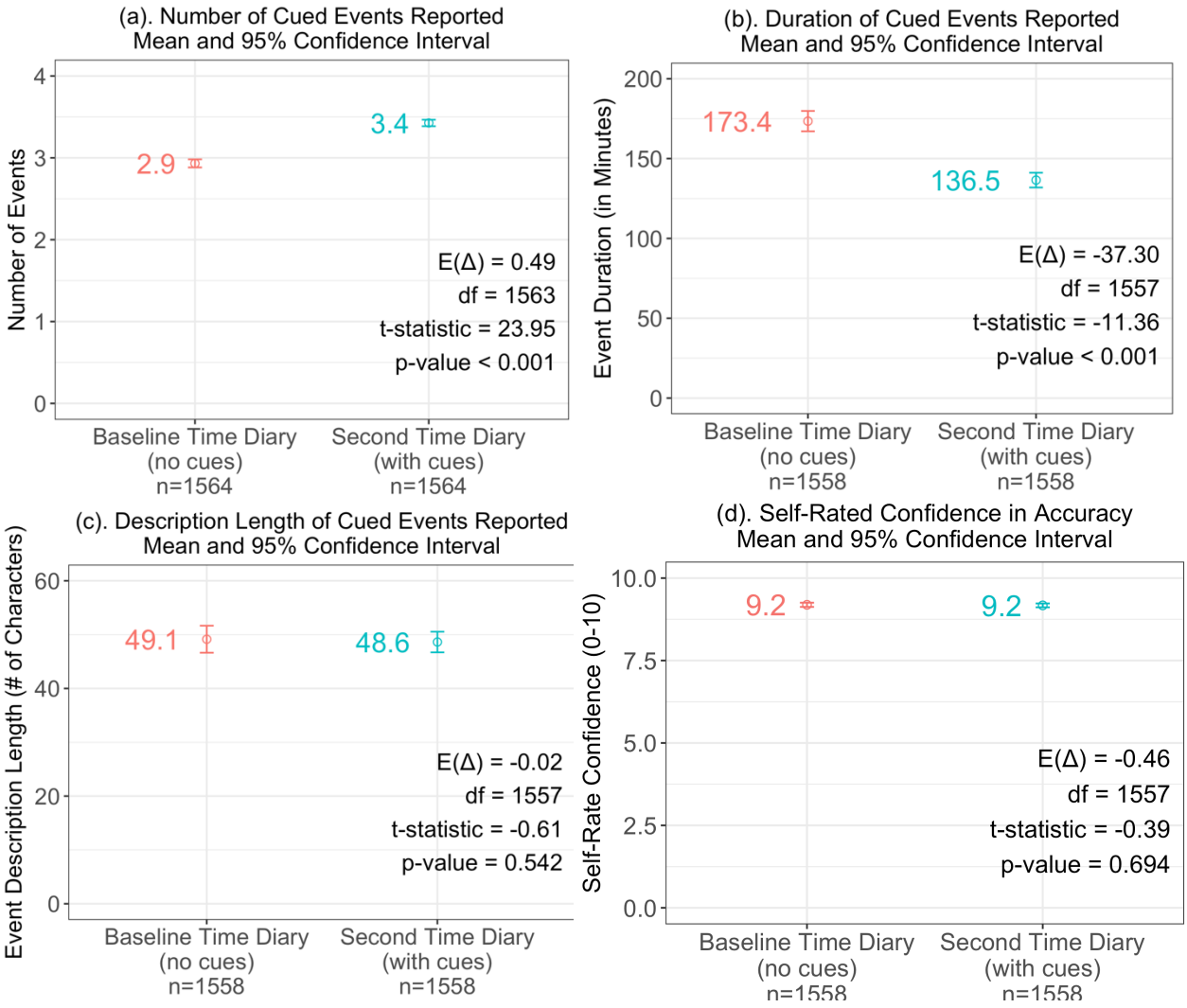


Figure 3.1 Data Quality Measures of Cued Events
(E(Δ) indicates the average difference between the baseline time diary and second time diary per respondent; t-statistics are based on a one-sample t-test of Δ; degree of freedom = n-1)

There were no overall differences in the average length of reported event descriptions or self-rated confidence in the accuracy of recall for cued events (see Figure 3.1c and 3.1d) when comparing cues versus no cues. The extremely positive confidence score for both time diaries suggests that respondents only reported events that they were reasonably sure of. The effect of the cues, therefore, was in fact to help respondents recall events that they were less sure of.

Moderators of Data Quality

We also examined if cue type affects recall on recall and if so, how the improvement varies among groups with different levels of memory capability (see additional tables in Table A3.1 – A3.10 for detailed models and tests).

Cue Type (H2)

Among the four quality measures are multiple pieces of evidence showing that map cues were not always as effective as other types of cues in facilitating recall when the location of an event was a respondent's home. Although there was no other evidence found that photo cues are more advantageous than other types of cues except that participants reported higher confidence about their recall when using photos. We found no differences between the two types of text cues (activity + location + participants vs. activity + feeling; RQ1); thus, no further discussion about the two text cues appears in this section.

Even though the map cue condition was found to promote recall of more cued events in the second time diary (see "Map cues" in Figure 3.3a), the kind of events whose recall it promoted were somewhat different from the kinds of events recalled in other cue conditions. More specifically, the map cue condition promoted significantly more false sleeping events[12] and

[12] Respondents should have always been awake when they replied to the text signals since the signals required human action using a mobile device. In other words, the cued events should very rarely be sleeping events (such as when a respondent had just gone to bed or woke up). A substantial increase in sleeping events in the second time diary raised a red flag about accuracy.

significantly fewer events such as “watching TV and playing video games” or “resting, waiting and doing nothing in particular” compared to other cue conditions. In summary, respondents were more likely to guess and infer what they were doing when the cue was a “home” location, especially when the signaled times were not close to their regular schedules (e.g., wake-up time, lunchtime, dinnertime). Map cues, meanwhile, resulted in no differences in other location-specific events for which location is unique and sufficient for recall. For instance, a normal commuting route is an effective cue for recalling “traveling”; a grocery store is effective for recalling “shopping.” Figure 3.2 shows the number of events of “watching TV and playing games” that all of the respondents reported and for which map cues were not as effective (the delta between the baseline and the second diary is significantly different from the average delta). The same comparison for other event types can be found in Figure A3.1 and Figure A3.2.

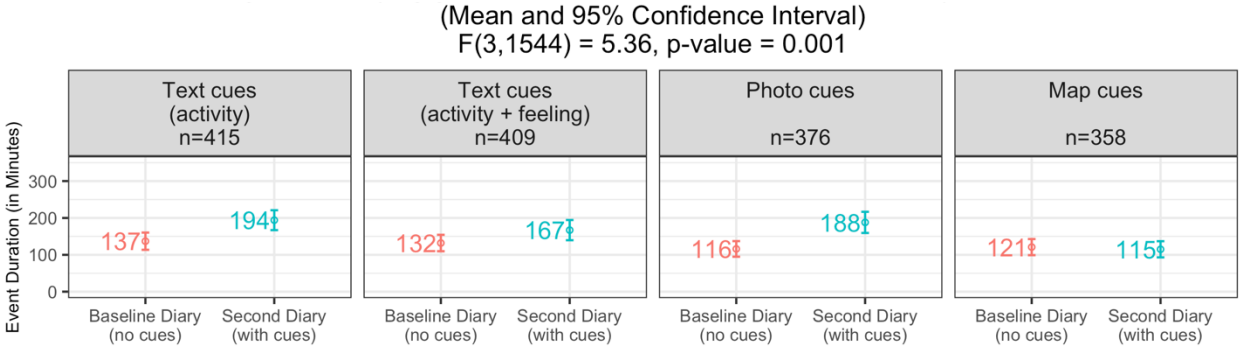


Figure 3.2 Number of Cued “Watching TV and Playing Games” Reported by All Respondents (ANOVA tests the difference between the two time diaries across the four conditions in a model that controls for age, gender, education level, day of the week, self-rated memory, and memory capability.)

The effect of map cues on event duration was also weaker than the effects of the other cue types (see Figure 3.3b). Figure 3.3 depicts the group means for the baseline and second time diaries by cue type. ANOVA tested the difference between the baseline and second time diaries across the four cue types. Controlling for demographic conditions, map cues reduced the average event duration by 12.2 minutes, but the effect was 31.59 minutes shorter than it was for the other cue conditions ($t(1546) = 4.26$, p-value <0.001; see Table A3.8 for details). This indicates that while map cues help with recall, they are not as effective as the other cue types for recalling smaller and more fine-grained events.

Presumably due to the limited information provided by a map, respondents tended to write shorter event descriptions in the second time diary than they did in the baseline time diary; their event descriptions were more or less the same length for both diaries in the other cue conditions (see Figure 3.3c). Controlling for demographic conditions, the map cue condition produced shorter event descriptions: 12 fewer characters in the map cue condition compared to other cue conditions ($t(1546) = -4.508$, p-value <0.001; see Table A3.9 for details). This suggests that map cues might be able to help individuals recall an event but do not provide too much detail about the event.

We hypothesized (Hypothesis 2) that photo cues have a stronger effect on recall because of the fine-grained, contextual information a photo can potentially include. We found partial support for this hypothesis: the subjective measure (i.e., self-rated confidence) showed a slight advantage when photo cues were presented compared to other type of cues (see Figure 3.3d). Controlling for demographic conditions, the average self-rated confidence score increased by 0.26

more from the baseline time diary to the second time diary (t(1546) = 3.343, p-value = 0.008; see Table A3.10 for details) for the photo cue condition compared to the other cue conditions.

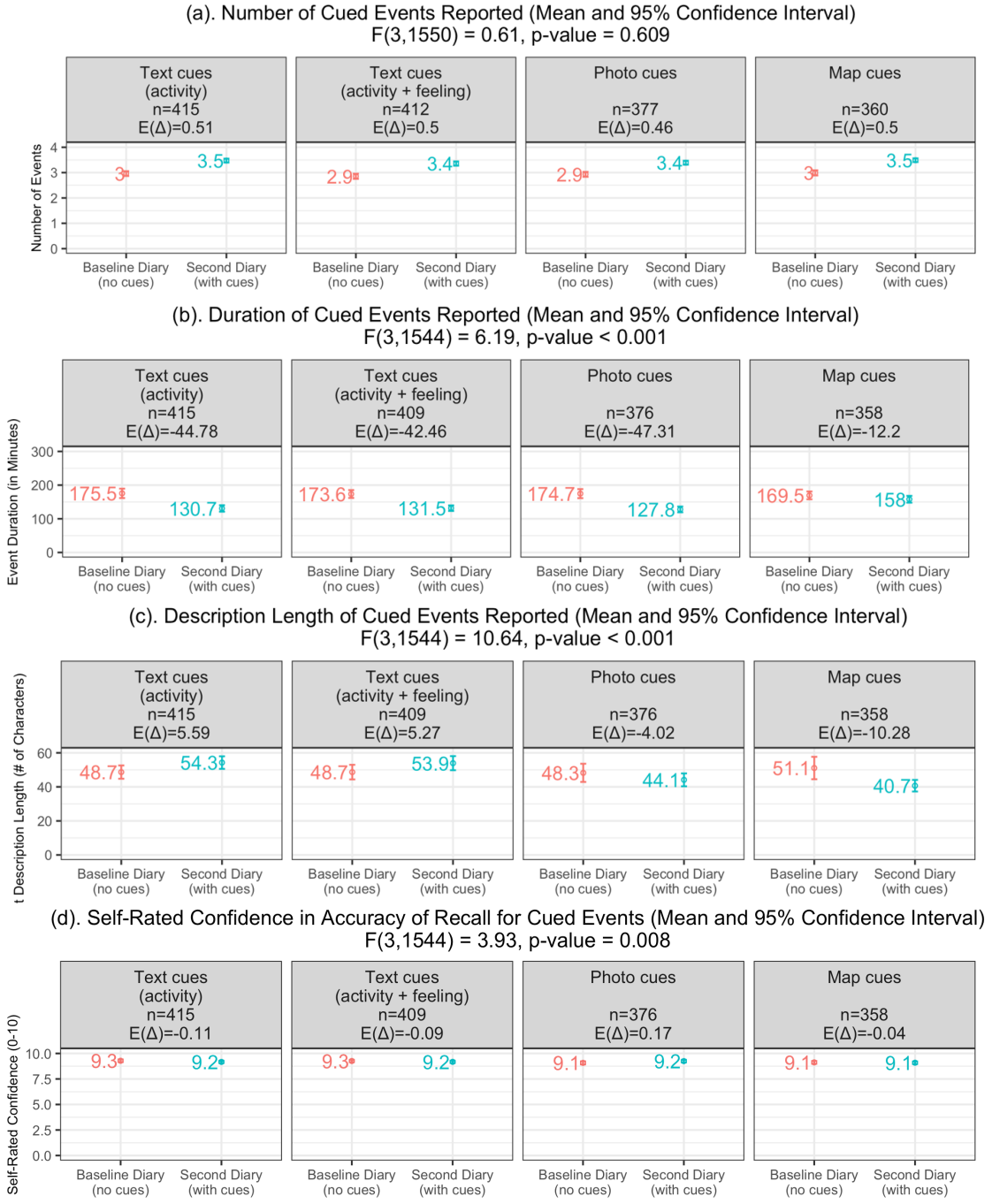


Figure 3.3 Moderating Effect of Cue Type on Data Quality Measures
(The F-test determines if the difference between the two time diaries are equal across the four conditions in Table A3.1 - A3.4, controlling for age, gender, education level, day of the week, self-rated memory, and memory capability.)

Memory Capability (H3)

We used the number of events reported in the baseline time diary as a proxy for memory capability. The cut-off we chose was >= 12 recalled events in the baseline time diary. We hypothesized that the effect of cues is stronger for individuals who have weaker memories because they have more room to improve. Their counterparts—individuals who can remember most of the events from the previous day—may perform near the ceiling, meaning that additional memory aids would not make a substantial difference. This hypothesis was supported by two measures of data quality. Individuals whose memory performance was relatively weak in the baseline diary recalled 0.83 more events in the second time diary. This represents a larger improvement (0.57 more cued events) than that which we observed for those with stronger memory performance ($t(1550)= 14.024$, $p<0.001$; see Table A3.7 for details).

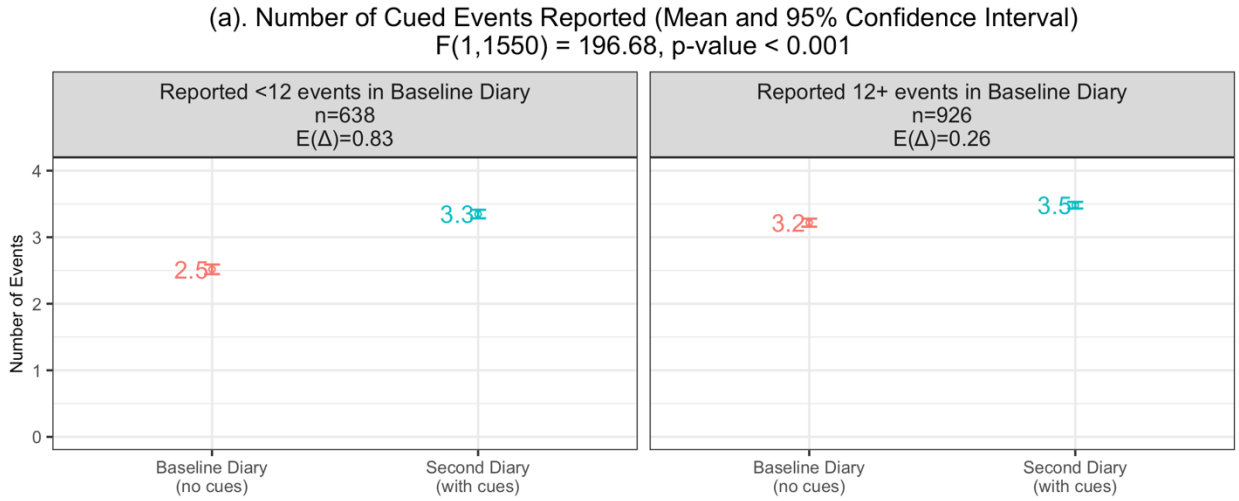


Figure 3.4 Number of Cued Events Reported by Memory Capability
(The F-test is based on the ANOVA model shown in Table A3.1, controlling for cue type, age, gender, education level, day of the week, and self-rated memory.)

In addition, the average event duration shows a similar pattern. There was a greater reduction in average event duration (more fine-grained) for participants who are relatively more forgetful: the average event duration was 84.09 minutes shorter than it was in the baseline time

diary, which is a reduction of 78.2 minutes more than was exhibited by the group with better memory (t(1550) = -12.017, p<0.001; see Table A3.8 for a detailed model).

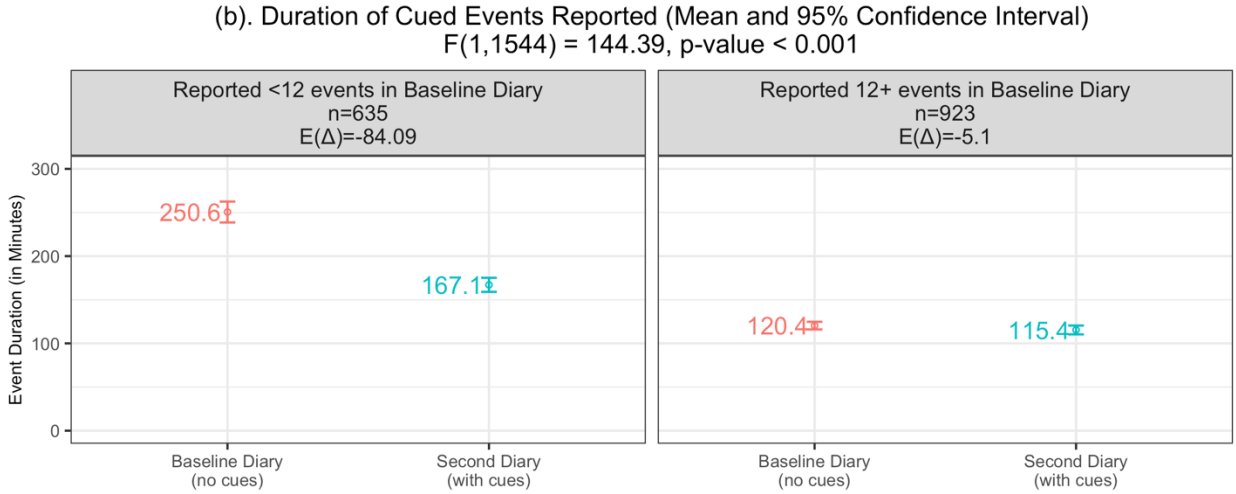


Figure 3.5 Average Event Duration of Cued Events by Memory Capability
(The F-test is based on the ANOVA model shown in Table A3.2, controlling for cue type, age, gender, education level, day of the week, and self-rated memory.)

Other Moderators

We found a marginal moderating effect for individuals with different education levels (F(1,1550) = 2.65, p-value = 0.071). The lower the level of education a participant had, the more events they recalled with the aid of the cues (see Figure A3.3), which means the participant benefited more from the cues. This trend makes sense considering that recall in the baseline time diary was lower for individuals who had no college degree, giving them more room to improve with the help of cues. Education usually corresponds to higher levels of cognitive capability and better memory capability, leaving less room for improvement with any memory aid.

The same logic suggests that older individuals are likely to benefit more from cues than are younger individuals: age is well known to be inversely correlated with memory ability (Craik and Jennings, 1992). We were unable to test this using our sample, however, because we simply had too few older participants. Apart from age, we found no difference in the effect of cues across

gender, day of the week on which events occurred, or self-rated memory. The sample size in this experiment was insufficient for further examining if and how different types of cues work better or worse for any particular demographic group.

Reporting Effort

Figure 3.6 Response Time per Cued Event (in Seconds)
(E(Δ) indicates the average difference between the baseline time diary and the second time diary per respondent; the t-statistics are based on a one-sample t-test of Δ; degree of freedom = n-1)

In completing the second time diary, respondents spent more time on processing, recalling, and reporting with the aid of the cues, which suggests the cues did motivate or facilitate additional effort in these recall tasks. On average, the respondents spent 16.9 more seconds per cued event in the second time diary ($t(1557) = 3.51$, $p < 0.001$).

At the end of each time diary, whether the baseline or the second time diary, respondents indicated how much effort they put into generating the cues on the previous day using the assigned format (text, photo, or map). Even though all ratings leaned toward the low extremes across cue types, the difference in perceived effort or burden between the baseline and second time diaries varied by cue type ($F(3,1540) = 6.94$, p-value <0.001, see Figure 3.7). Not surprisingly, recording one's GPS location was perceived to be easier than providing a text description or taking a photo.

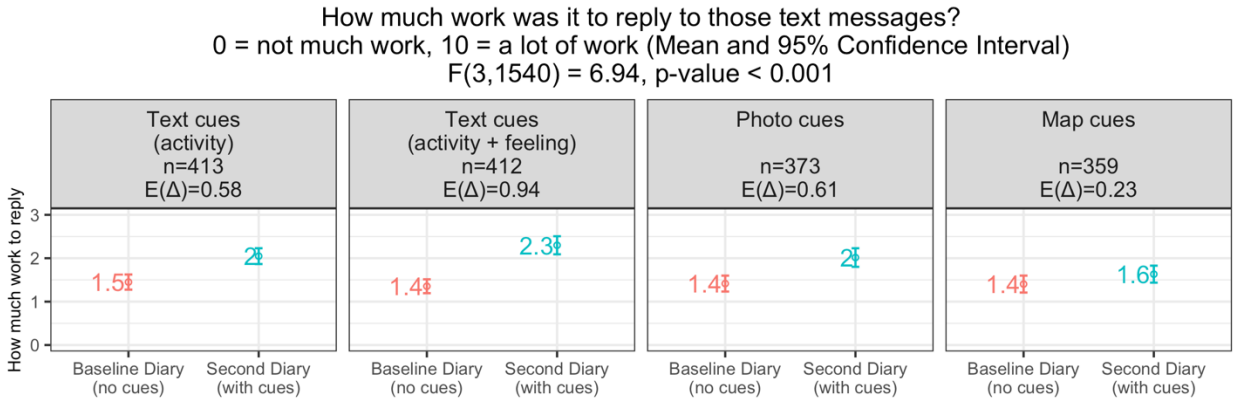


Figure 3.7 Perceived Burden of Replying to Text Messages by Formality
(F-test is based on the ANOVA model shown in Table A3.6, controlling for cue type, age, gender, education level, day of the week, and self-rated memory.)

3.5 Discussion

This study demonstrated the feasibility of collecting textual, photographic, and geographical event descriptions from individuals using their smartphones in a naturalistic survey setting and using the collected descriptions as memory cues to facilitate recall in a customized follow-up web survey. It also demonstrated that such memory cues help individuals remember and report more events than they are able to during spontaneous recall without cues. Reported events with cues were found to be more fine-grained than those without cues regardless of the cue type (either kind of text, photo, or map).

There is a trade-off between the quality of the data when cues are presented at the time of recall and the effort required of individuals to create the cues (event descriptions) during the day about which they will later be asked to report. Map cues require the least effort but might provide the least information for recall compared to the other types of cues (the two types of text cues and photo cues). Whether a map cue suffices very much depends on: (1) the mobility of an individual over the course of the day, (2) the uniqueness of the location where the event took place, and (3) the purpose of the study (the granularity required by the study). Map cues are effective to the extent

that a study does not require a granular self-report of events and that the participant frequently moves to different locations throughout the day and does different things at each location. For instance, a survey about transportation such as the National Household Travel Survey (Santos et al., 2011) asks individuals to recall every single travel episode on an actual workday or a weekend day. Various locations (shown on a map) at different times of the day would be effective cues for this survey's respondents as they attempt to recall every travel episode on the given day and the means of transportation for each. Map cues would likely not be as effective as text or photo cues in surveys that measure only at-home activities since map cues do not provide any additional information to distinguish events that occur at a single location from one another. We found only some evidence from the participants' subjective confidence ratings to support our prediction that photo cues might be more advantageous than other types of cues. Future studies are needed to fully demonstrate the potential of photo cues (possibly in a laboratory setting or researcher-administered interviews), followed by additional efforts to determine how to take advantage of photo cues and make them more effective in naturalistic survey settings in which respondents administer the surveys themselves.

This study also provides evidence that the individuals who would most benefit from the MB approach are those whose memory capability is relatively weak. These individuals have much more room to improve since they are unable to remember and report many events in the absence of cues. Their counterparts with greater recall capability can remember and report most events without a memory aid and thus would benefit less. There are potential moderating characteristics such as age that we were unable to fully test in this experiment. However, based on the current findings regarding memory capability and education, we hypothesize that the MB method would work better (i.e., improve memory more) for older individuals than for younger individuals.

In this study, we did not find that additional mood information improved later recall when the information was presented as part of a text cue. There were no differences between the two types of text cues across all measures. This suggests that either that the activity description in both types of text cues was too strong for other context (including location, participants, and affect [mood]) to matter or that the affect information was as effective as the combined information about the location and participants.

Here, we would like to clarify how the proposed approach is different from the ecological momentary assessment (EMA) and the experience sampling method (ESM) and to explain when the proposed approach might be better. The EMA (Stone and Shiffman, 1994) and ESM (Csikszentmihalyi et al., 1977) approaches—which are virtually the same—usually require respondents to answer a series of questions *in situ*. For instance, in the first ESM study, Csikszentmihalyi and his colleagues (1977) asked about 21 questions at every signaled moment and signaled respondents about 5–7 times per day. Such intensive data collection is burdensome and disrupts individuals' daily activities. It may also affect how individuals behave as they become more aware of the behaviors they report (i.e., reactivity). The MB approach can be seen as an extended version of the traditional EMA/ESM technique, in which we try to reduce the extent to which the collection of cues disrupts respondents' daily activities and affects their subsequent behavior. We require only one text, one photo, or one geolocation at the signaled times and use the collected *in situ* information as memory cues *after the fact* rather than treating it as *the data*. In this experiment, we showed that these cues are both useful (i.e., they promote recall) and a low burden on respondents.

3.6 Limitations

Validating self-reported events against actual events was extremely hard in our naturalistic survey setting, given that we had no records of participants' personal events. Comparing the participants' reports across two time diaries, where cues either were or were not presented in the second time diary, served as a proxy for recall accuracy given the naturalistic setting in which the study was conducted and the consequential lack of validation data.

In this study, we used a convenience sample rather than a probability-based sample due to financial constraints. However, because we balanced the key demographics of gender, age, education level, day of the week, self-perceived memory, and memory capability. We also used repeated measures design and thus the results of the study are reasonably generalizable to the broader population. Nevertheless, we had very little control over the overall demographics. As a result, we were not able to test the extent to which the MB approach helped all demographic subpopulations, such as older individuals. It would be worthwhile to test the method for older individuals in particular, as the results might help advance research on older adults, such as that which comprises the Health and Retirement Study.

References

Blaney, P. H. (1986). Affect and memory: a review. *Psychological bulletin*,*99*(2), 229.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325-14329.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In Neisser, U. and Winograd, E., In Neisser, U., & Winograd, E. (Eds.). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*, pp. 21-90. Cambridge University Press, New York.

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, *52*(4), 618-627.

Craik, F. I. M., & Jennings, J. M. (1992). Human memory. In E. Fergus, I. M., Craik, E. Timothy, & A. Salthouse (Eds.), *The handbook of aging and cognition* (pp. 51–110). Hillsdale, NJ: Erlbaum.

Csikszentmihalyi, M., Larson, R., & Prescott,S. (1977). The ecology of adolescent activity and experience. *Journal of youth and adolescence, 6*(3), 281-294.

Eich, E., Macaulay, D., & Ryan, L. (1994). Mood dependent memory for events of the personal past. *Journal of Experimental Psychology: General*,*123*(2), 201.

Fisher, K., & Gershuny, J. (2013). *Time use and time diary research*. Oxford University Press.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., ... & Wood, K. (2006, September). SenseCam: A retrospective memory aid. In *International Conference on Ubiquitous Computing* (pp. 177-193). Springer, Berlin, Heidelberg.

Kelly, P., Thomas, E., Doherty, A., Harms, T., Burke, Ó., Gershuny, J., & Foster, C. (2015). Developing a method to test the validity of 24 hour time use diaries using wearable cameras: a feasibility pilot. *PLoS One*, *10*(12), e0142198.

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics*, *80*(5), 1035-1056.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, *6*(1), 156-163.

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*.

Rubin, D. C., Schrauf, R. W., & Greenberg, D. L. (2003). Belief and recollection of autobiographical memories. *Memory & cognition*, *31*(6), 887-901.

Tulving, E. (1974). Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist*, 62(1):74–82.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352–373.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. Cognitive psychology, 18(2):225–252.

Yan, T. & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1):51–68.

Appendices

Appendix 3.1 Additional Tables

Table A3.1 ANOVA Model of Difference in Number of Reported Cued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	1.073	3	0.610	0.609
Age	3.607	3	2.050	0.105
Gender	0.017	1	0.028	0.867
Highest Education Degree	3.104	2	2.645	0.071
Day of the Week	0.963	1	1.641	0.200
Self-rated Memory	1.290	2	1.100	0.333
Memory Capability	115.387	1	196.684***	<0.001
Residual	909.326	1550	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.

Table A3.2 ANOVA Model of Difference in Average Event Duration of Cued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	281462.513	3	6.194***	<0.001
Age	93907.045	3	2.067	0.103
Gender	8560.609	1	0.565	0.452
Highest Education Degree	17478.465	2	0.577	0.562
Day of the Week	17651.290	1	1.165	0.281
Self-rated Memory	5515.559	2	0.182	0.834
Memory Capability	2187062.013	1	144.386***	<0.001
Residual	23387521.288	1544	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.3 ANOVA Model of Difference in Average Event Description Length of Cued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	66516.940	3	10.642***	<0.001
Age	10143.390	3	1.623	0.182
Gender	2557.870	1	1.228	0.268
Highest Education Degree	14296.162	2	3.431*	0.033
Day of the Week	2135.577	1	1.025	0.311
Self-rated Memory	73.177	2	0.018	0.983
Memory Capability	7397.080	1	3.550	0.060
Residual	3216863.710	1544	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.4 ANOVA Model of Difference in Average Self-Rated Confidence of Cued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	19.749	3	3.934**	0.008
Age	6.067	3	1.208	0.305
Gender	1.871	1	1.118	0.291
Highest Education Degree	3.149	2	0.941	0.391
Day of the Week	0.019	1	0.011	0.915
Self-rated Memory	0.630	2	0.188	0.828
Memory Capability	0.008	1	0.005	0.945
Residual	2584.006	1544	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.5 ANOVA Model of Difference in Response Time per Cued Event Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	146282.747	3	1.354	0.255
Age	224879.077	3	2.082	0.101
Gender	1353.654	1	0.038	0.846
Highest Education Degree	88418.693	2	1.228	0.293
Day of the Week	94.515	1	0.003	0.959
Self-rated Memory	168603.806	2	2.341	0.097
Memory Capability	86943.204	1	2.415	0.120
Residual	55592785.788	1544	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.6 ANOVA Model of Difference in Perceived Burden of Replying to Text Messages Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	101.814	3	6.936***	<0.001
Age	24.035	3	1.637	0.179
Gender	0.049	1	0.010	0.920
Highest Education Degree	4.472	2	0.457	0.633
Day of the Week	10.496	1	2.145	0.143
Self-rated Memory	10.711	2	1.095	0.335
Memory Capability	0.020	1	0.004	0.949
Residual	7535.207	1540	--	--

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.7 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on the Number of Cued Events

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	0.149	0.087	1.707	0.088
Cue Type				
Text Cue (Activity only)	--	--	--	--
Text Cue (Activity + Mood)	-0.031	0.053	-0.590	0.555
Photo Cue	-0.072	0.055	-1.311	0.190
Map Cue	-0.018	0.055	-0.325	0.745
Age				
18–24 years old	--	--	--	--
25–34 years old	0.035	0.054	0.645	0.519
35–44 years old	0.136*	0.060	2.246	0.025
45+ years old	0.070	0.071	0.991	0.322
Gender				
Male	--	--	--	--
Female	-0.007	0.042	-0.168	0.867
Highest Education Degree				
Graduate degree	--	--	--	--
College degree	0.102	0.054	1.865	0.062
No college degree	0.122*	0.054	2.253	0.024
Day of the Week				
Weekday	--	--	--	--
Weekend	-0.050	0.039	-1.281	0.200
Self-rated Memory				
Excellent	--	--	--	--
Very good	0.057	0.046	1.222	0.222
Good/Fair/Poor	-0.002	0.055	-0.038	0.970
Memory Capability				
12+ events in Baseline	--	--	--	--
<12 events in Baseline	0.567***	0.040	14.024	<0.001

Note: Dependent Variable = Second Time Diary – Baseline Time Diary, N = 1564, Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A3.8 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Event Duration

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	-2.507	13.097	-0.191	0.848
Cue Type				
Non-map cue	--	--	--	--
Map cue	31.590***	7.422	4.26	<0.001
Age				
18–24 years old	--	--	--	--
25–34 years old	-11.763	8.743	-1.35	0.179
35–44 years old	-23.924*	9.724	-2.46	0.014
45+ years old	-11.317	11.414	-0.99	0.322
Gender				
Male	--	--	--	--
Female	5.077	6.792	0.748	0.455
Highest Education Degree				
Graduate degree	--	--	--	--
College degree	-2.506	8.763	-0.286	0.775
No college degree	-8.274	8.743	-0.946	0.344
Day of the Week				
Weekday	--	--	--	--
Weekend	6.855	6.327	1.083	0.279
Self-rated Memory				
Excellent	--	--	--	--
Very good	-2.100	7.476	-0.281	0.779
Good/Fair/Poor	2.595	8.879	0.292	0.770
Memory Capability				
12+ events in Baseline	--	--	--	--
<12 events in Baseline	-78.205***	6.508	-12.017	<0.001

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
N = 1558
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.

Table A3.9 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Event Description Length

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	9.835	4.875	2.018	0.044
Cue Type				
Non-map cue	--	--	--	--
Map cue	-12.452***	2.762	-4.508	<0.001
Age				
18–24 years old	--	--	--	--
25–34 years old	-2.807	3.254	-0.863	0.388
35–44 years old	-1.750	3.619	-0.484	0.629
45+ years old	5.061	4.248	1.191	0.234
Gender				
Male	--	--	--	--
Female	-2.584	2.528	-1.022	0.307
Highest Education Degree				
Graduate degree	--	--	--	--
College degree	-2.333	3.261	-0.715	0.475
No college degree	-7.337*	3.254	-2.255	0.024
Day of the Week				
Weekday	--	--	--	--
Weekend	2.403	2.355	1.020	0.308
Self-rated Memory				
Excellent	--	--	--	--
Very good	0.106	2.782	0.038	0.970
Good/Fair/Poor	-0.472	3.305	-0.143	0.886
Memory Capability				
12+ events in Baseline	--	--	--	--
<12 events in Baseline	-4.656	2.422	-1.922	0.055

Note: Dependent Variable = Second Time Diary – Baseline Time Diary,N = 1558
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.

Table A3.10 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on Average Self-Rated Confidence Score of the Reports

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	-0.230	0.139	-1.654	0.098
Cue Type				
Non-photo cue	--	--	--	--
Photo cue	0.256**	0.077	3.343	0.001
Age				
18–24 years old	--	--	--	--
25–34 years old	0.094	0.092	1.027	0.304
35–44 years old	-0.057	0.102	-0.555	0.579
45+ years old	-0.002	0.120	-0.013	0.990
Gender				
Male	--	--	--	--
Female	0.075	0.071	1.046	0.296
Highest Education Degree				
Graduate degree	--	--	--	--
College degree	0.125	0.092	1.362	0.173
No college degree	0.097	0.092	1.051	0.293
Day of the Week				
Weekday	--	--	--	--
Weekend	0.007	0.067	0.104	0.917
Self-rated Memory				
Excellent	--	--	--	--
Very good	-0.045	0.079	-0.573	0.566
Good/Fair/Poor	-0.013	0.093	-0.140	0.889
Memory Capability				
12+ events in Baseline	--	--	--	--
<12 events in Baseline	0.004	0.068	0.066	0.948

Note: Dependent Variable = Second Time Diary – Baseline Time Diary, N = 1558
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Appendix 3.2 Additional Figures

Figure A3.1 At-home Events for Which Map Cues Are Not as Effective as Other Cues (ANOVA tests the difference between the two time diaries across the four conditions in a model that controls for age, gender, education level, day of the week, self-rated memory, and memory capability.)

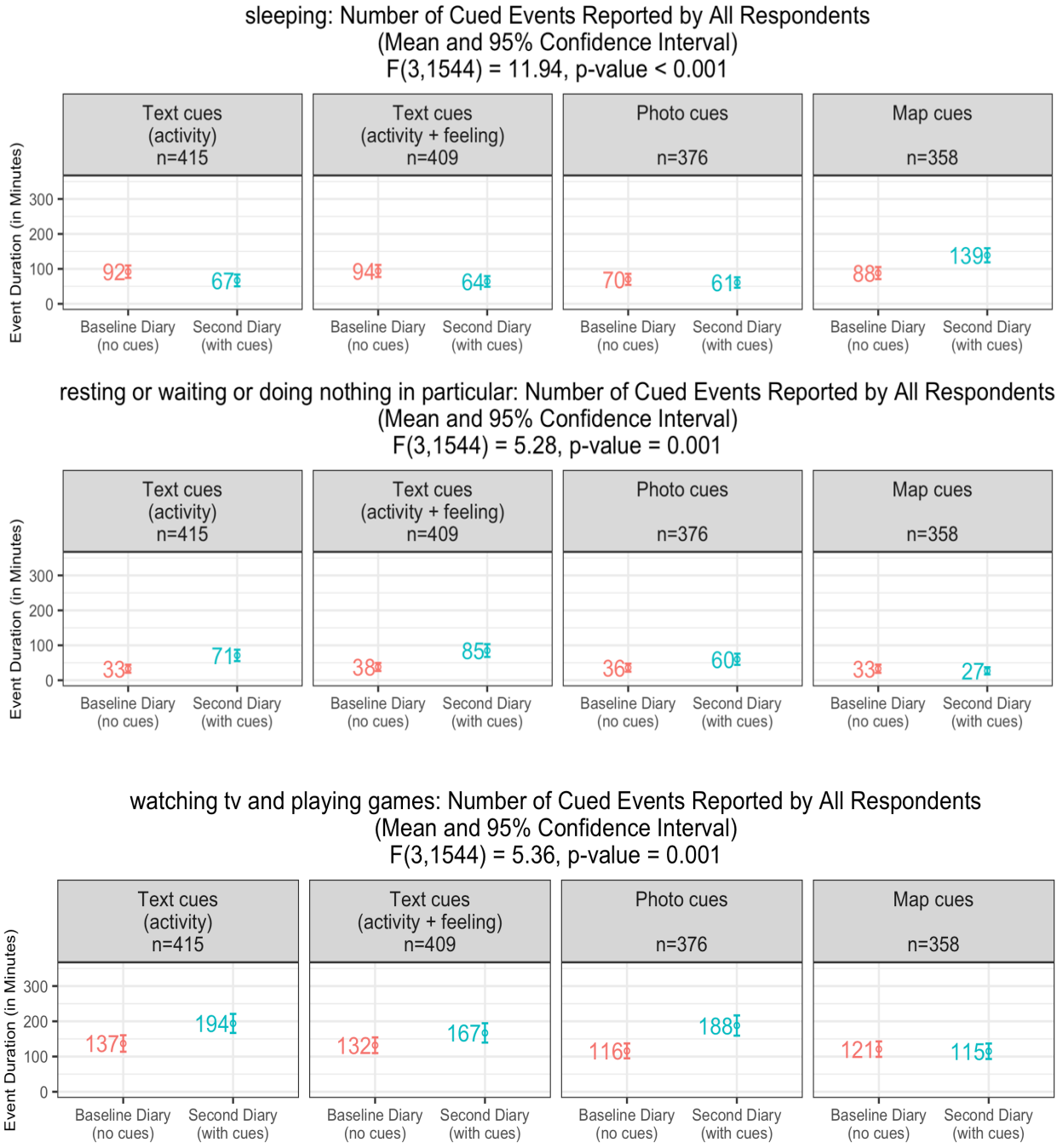
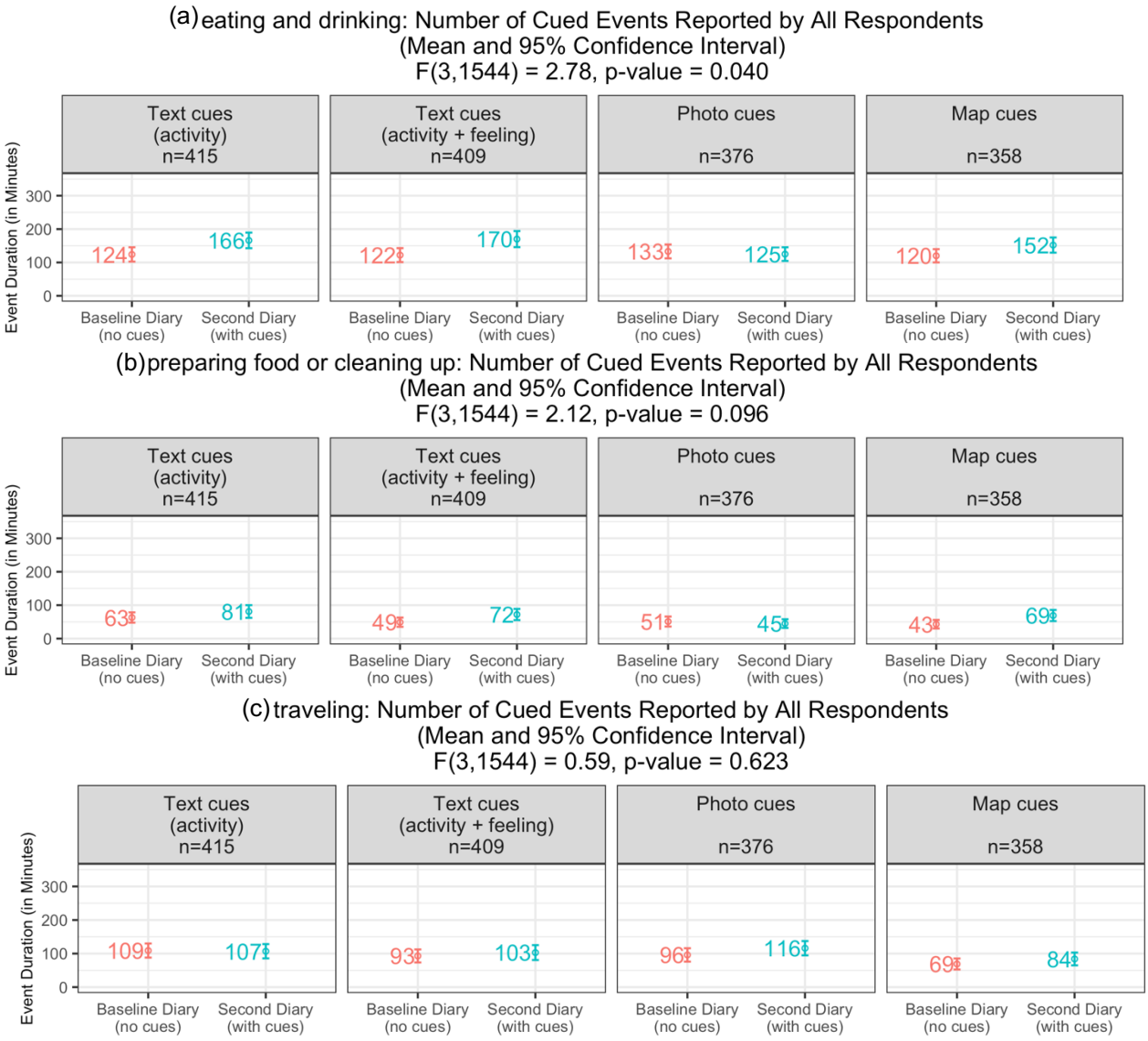
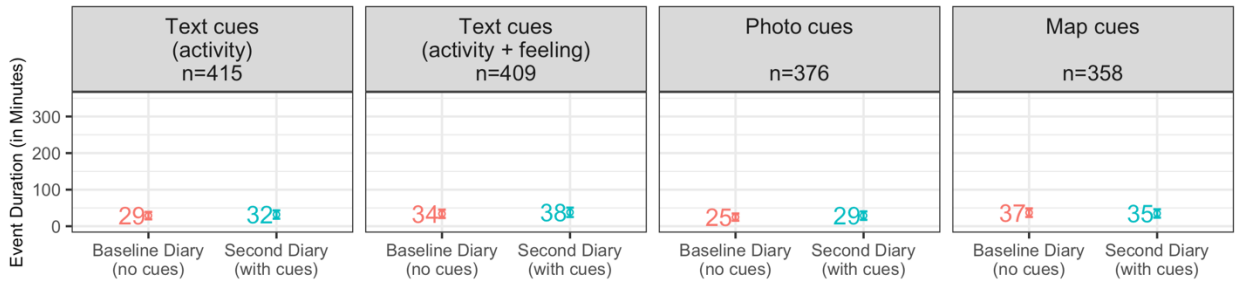


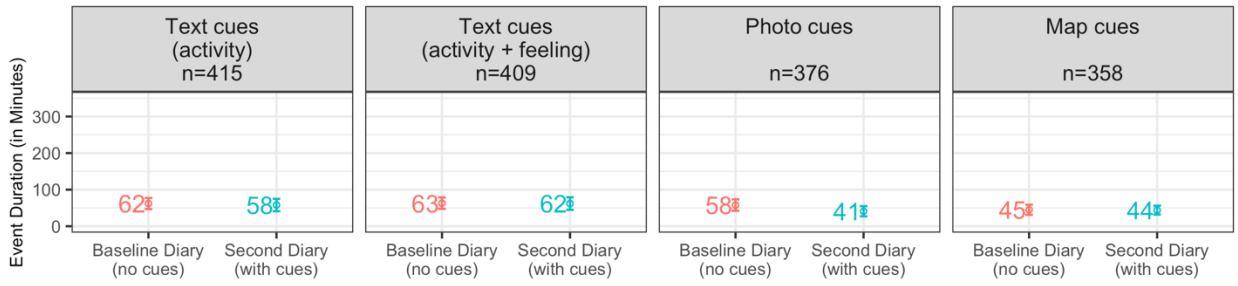
Figure A3.2 Events for Which Map Cues Are Similar to Other Cues
(ANOVA tests the difference between the two time diaries across the four conditions in a model that controls for age, gender, education level, day of the week, self-rated memory, and memory capability.)



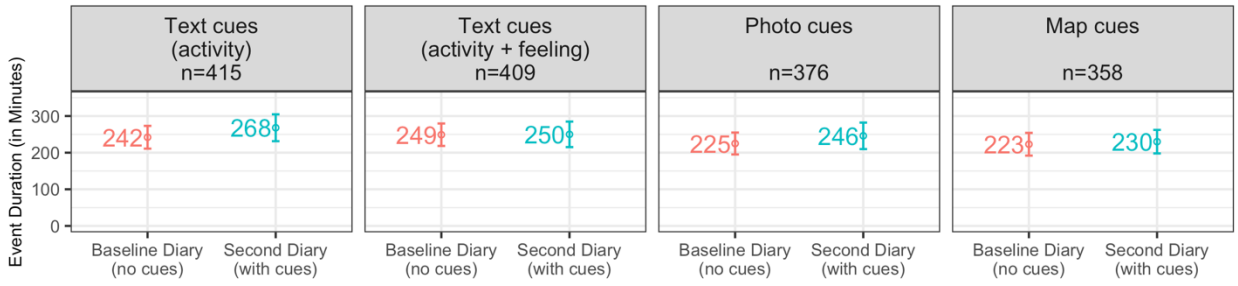
(d) doing sports or exercising or walking: Number of Cued Events Reported by All Respondents
(Mean and 95% Confidence Interval)
F(3,1544) = 0.16, p-value = 0.926



(e) socializing with others or phone calls or texting: Number of Cued Events Reported by All Respondents
(Mean and 95% Confidence Interval)
F(3,1544) = 0.74, p-value = 0.529



(e) working or studying: Number of Cued Events Reported by All Respondents
(Mean and 95% Confidence Interval)
F(3,1544) = 0.52, p-value = 0.665



(f) shopping including groceries: Number of Cued Events Reported by All Respondents
(Mean and 95% Confidence Interval)
F(3,1544) = 0.49, p-value = 0.690

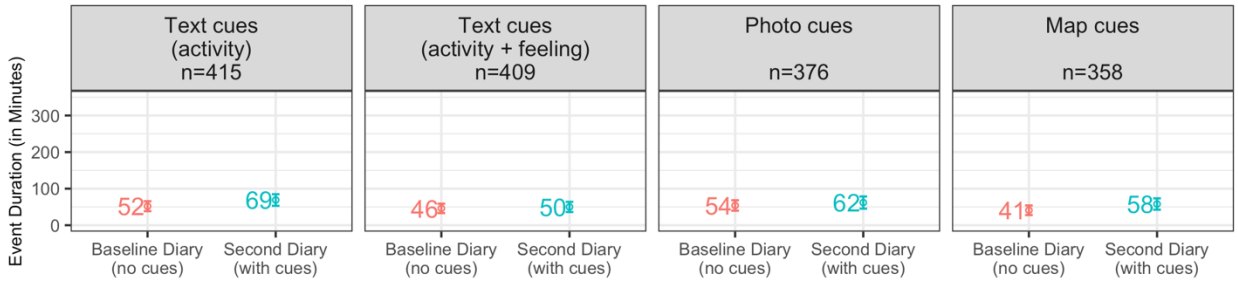
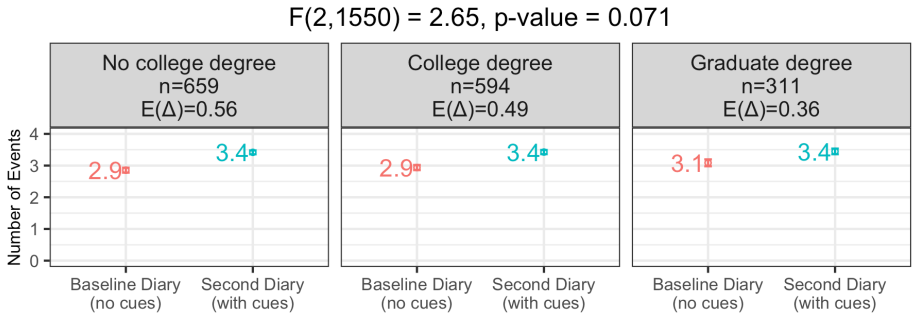


Figure A3.3 Number of Cued Events by Education Level
(ANOVA tests the difference between the two time diaries across the four conditions in a model that controls for age, gender, education level, day of the week, self-rated memory, and memory capability.)



Chapter 4 Using Text, Photos, and Maps as Recall Cues in a Time Diary on the Web

4.1 Introduction

Time diaries are used and collected across the world. There are 85 countries worldwide that have conducted time diary studies between 1966 and 2015 (United Nations). Labor economic studies rely on time-use data or time-budget data to analyze issues related to time allocation in the labor market (see Hamermesh and Pfann, 2005, for a review). In addition to containing information about the actual activities that occurred on a particular day, time diaries often measure individuals' well-being. For instance, the subjective well-being module in the American Time Use Survey provides a basis for understanding the well-being of the nation, or individuals' subjective satisfaction from work, leisure, and other activities (Stone and Mackie, 2014).

Accurate recall and good-quality data are crucial in order to make accurate inferences about a population. Time diaries require respondents to accurately recall from episodic memory a full sequence of a recent day or two. The traditional time diary approach relies on purely retrospective recall via various modes of data collection, including paper and pencil (e.g., Statistics Bureau Japan, 2016), telephone interviews (American Time Use Survey [ATUS], Bureau of Labor Statistics, 2016), the web (e.g., Statistics Canada, 2016), and mobile apps (Fernee et al., 2013; Chatzitheochari et al., 2015). Regardless of which mode is used to collect time diary information, time diaries are quite a burdensome recall task for individuals to complete. Thus far, there have been very few studies about helping respondents remember all of the personal events that occurred on a given day.

Two pilot studies discussed in Chapter 3 (Hodges et al., 2006; Kelly et al., 2015) introduced a rather intrusive method requiring subjects to carry a wearable camera for a day or more to record all of their activity. The photos captured are later used to prompt the subjects to provide further details about those moments. Considering that such a method probably cannot be used for a more general population and for every kind of study setting, it is important to develop alternative methods of helping individuals with the recall required by time diaries.

4.2 This Study

In the previous chapter, we evaluated the effect of cues on recalling cued events. This chapter focuses on whether cues would affect individuals recalling uncued events in a time diary.

The rationale for why cues might help respondents remember uncued events is that all the events that occur in a single day must have temporal and potentially causal relationships with each other. The cued events, once recalled, can serve as related anchors that can help respondents remember or reconstruct what was happening before and after those cued events. This is similar to event history calendar (EHC) approach, but the effectiveness of the cues might vary substantially. Unlike meaningful landmarks of a person's life in the EHC, the cues collected using the MB approach are random events captured during the day.

In this study, we again test four different types of recall cues to see if they promote recall of uncued events better than no cues: a textual description of the event (the activity, participants, and location); a textual description of the event and the respondent's mood at the time of the event; a photo taken by the respondent during the event; and a map showing the location of the event based on geocoordinates recorded by the respondent's smartphone. Secondly, we look at the time diary as a whole and evaluate whether the proposed memory bookmarking (MB) approach produces a net benefit for time diary accuracy and completeness. Given there are no prior studies

that address this issue, we lack strong hypotheses. Instead, we are exploring the answers to the following research questions:

Research question 1. Compared to spontaneous recall (free recall), do recall cues based on in situ verbal descriptions of activities, photos, and location information promote better recall of uncued personal activities captured in a time diary?

Research question 2. Do different cue types have different effects on the recall of uncued events?

Research question 3. Does recall cues based on in situ information have different effect on the recall of uncued events for individuals who have low memory capability ?

Research question 4. Does the proposed MB approach provide an overall net benefit for the completion of the time diary?

Research question 5. Does the proposed MB approach make the overall time diary task more burdensome since the MB approach entails two phases of data collection?

4.3 Methods

There were five experimental conditions with approximately 400 respondents in each: (1) *Control*, with no cues for either time diary; (2) *Text cue (activity, location and participants)*, which contains information about the activity, location where it occurred, and participants involved; (3) *Text cue (activity and mood description)*, which contains information about the activity and the respondent's mood at the moment; (4) *Photo*; and (5) *Map.* Randomization was balanced across the levels of six blocking factors: day type (weekday or weekend day), gender, age, education level, self-assessed memory, and the number of events reported in the baseline time diary (<12 events vs. 12+ events). A detailed description and demonstration of the *Experimental Design* can be found in Chapter 2.

Measures of Data Quality and Response Burden

We compared data quality and respondent effort between the two time diaries and across experimental groups using the indicators shown in Table 1 of Chapter 3; we also used the additional measures shown in the following table. As described in the *Experimental Design* in Chapter 2, we collected two additional GPS coordinates on the day before the time diary began from each of the respondents in both waves. The GPS coordinates were shown to the respondents in Google Maps after the respondents had completed their time diaries. The respondents were asked to label what each location meant, e.g., whether it is their home or yard, workplace, or school. We used these labels with their associated timestamps as location benchmarks (ground truth) to compare to the locations reported in the time diaries. The agreement between the two is a dichotomous variable that indicates the accuracy of the locations reported in the respondents' time diaries, with 1 suggesting that a location benchmark was correctly reported in a time diary and 0 suggesting that the location was incorrectly or not at all reported in a time diary.

Table 4.1 Additional Measures of Data Quality for the Time Diary as a Whole

Measure	**Type**	**Indication**	**Statistical test**	**Relevant literature**
Temporal gap	Continuous	Shorter temporal gap between reported activities means more complete time diary.	1. **Main effect** of the cues: - T-test for testing difference (Δ) between baseline and second time Diaries. 2. **Moderating effect** of the cue type and other factors: - Dependent variable is the difference (Δ). - Independent variable includes moderators and other blocking factors. - ANOVA - Parameter estimate testing in linear regression.	Fisher and Gershuny, 2013
Accuracy of reported location (Agreement between GPS location and reported location in the time diary)	Dichotomous	With 1 being accurate and 0 being inaccurate or missing.		Elevelt et al., 2019

Analysis Approach

When assessing the effect of cues on uncued events as well as on the entire time diary, we compared all of the selected measures between the two time diaries for MB conditions only and tested the difference (Δ) between the two time diaries by cue type and other blocking factors. We also compared the difference (Δ) in the two time diaries between the control condition and MB conditions for the overall time diary.

Note that in Chapter 3, we used "cued events" to indicate the signaled events in the second time diary and their temporal counterparts in the baseline time diary. In this chapter, we use "uncued events" to indicate events reported in either of the time diaries that occurred at uncued times in the second time diary. R was used for data management, statistical testing, and the modeling and generation of the figures.

4.4 Results

Effect on Data Quality of Uncued Events Among MB Conditions (RQ1)

Combining all four MB conditions, Figure 4.1 shows group means for the four quality measures for both time diaries. In summary, the respondents reported 1.25 fewer uncued events in the second time diary ($t(1563) = -9.61$, $p < 0.001$; see Figure 4.1a); the average event duration was 28.6 minutes longer ($t(1559) = 15.68$, $p < 0.001$; see Figure 4.1b); the length of the event description was 5.5 characters shorter ($t(1558) = -8.56$, $p < 0.001$; see Figure 4.1c); and there was no significant change in perceived confidence about the events reported (see Figure 4.1d). The effect on uncued events contrasts with that on cued events (see Figure 3.1 in Chapter 3). That is, presenting memory cues in a time diary seems to produce a positive effect on the recall of cued events (Figure 3.1, Chapter 3), while it produces no effect or even a negative effect on the recall

and/or reporting of uncued events (Figure 4.1). Looking at the benefits of the MB approach for cued and uncued events suggests a zero-sum game.

Respondents did report that memory cues were somewhat helpful (i.e., "*When you filled out your diary, how much did reading the memory cues help you remember OTHER activities? 0 = did not help much, 10 = helped a lot*"). The 95% confidence interval of the average rating was 6.2 to 6.6. We will talk about the misalignment between subjective and objective metrics in the *Discussion* section.

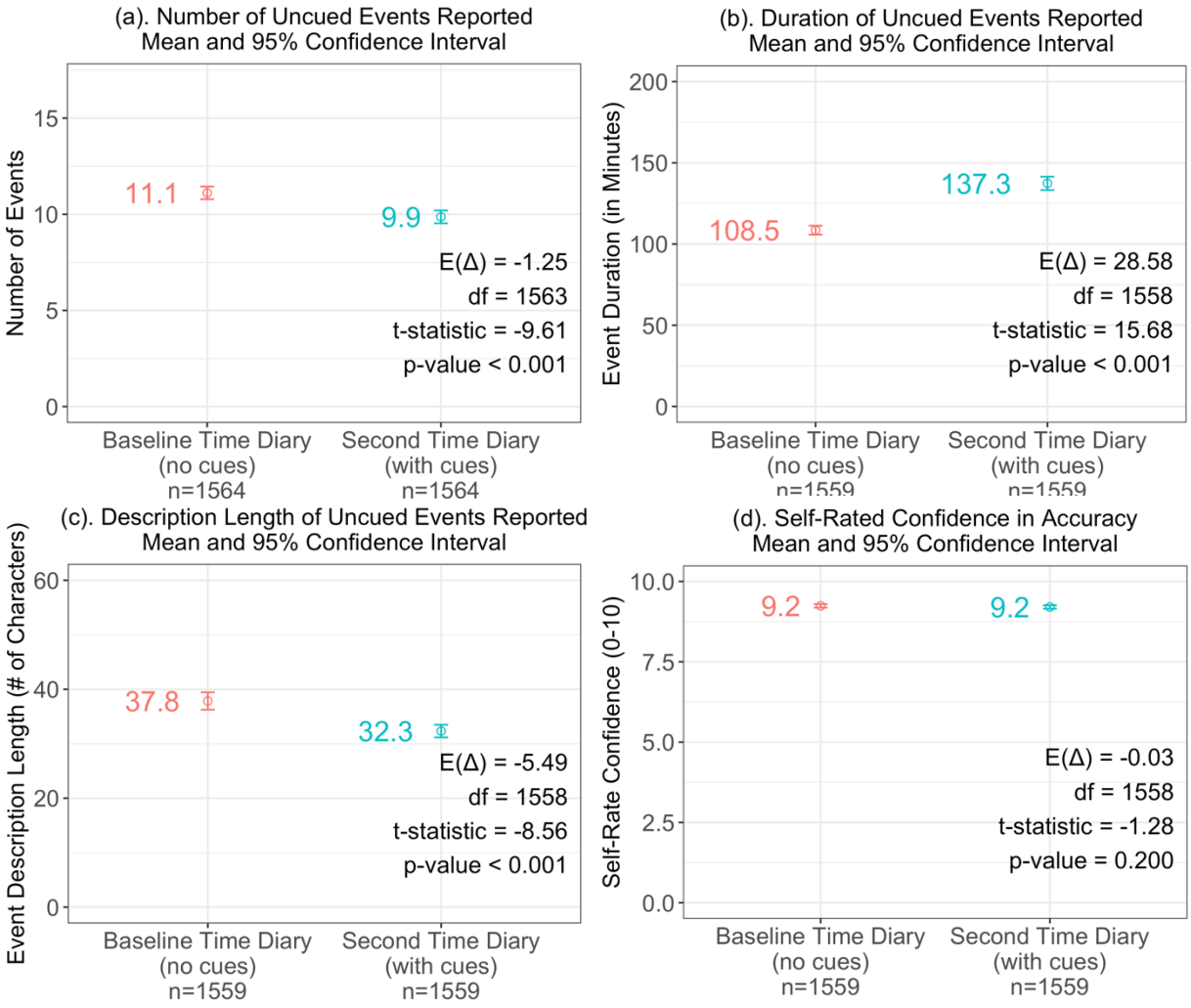


Figure 4.1 Data Quality Measures of Uncued Events
(E(Δ) indicates the average difference between the baseline time diary and second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1)

Cue Type (RQ2)

Breaking down these measures by cue type, we found that cue type had no impact on the number of uncued events, event duration, or perceived confidence rating (see Table A4.1, A4.2, A4.4 for detailed ANOVA models and F tests). Cue type only had an impact on the length of the textual description, which is consistent with its effect on cued events (see Figure 3.3, Chapter 3). However, the two types of text cues did not lead to longer descriptions in the second time diary, just less short descriptions compared to the other non-text cues (see Table A4.3 and Figure A4.1c for detailed models and tests). Respondents did assign a lower subjective rating[13] to map cues (see Figure 4.2 below; $F(3,1108) = 8.49$, $p < 0.001$) when assessing how much cues helped them remember other events.

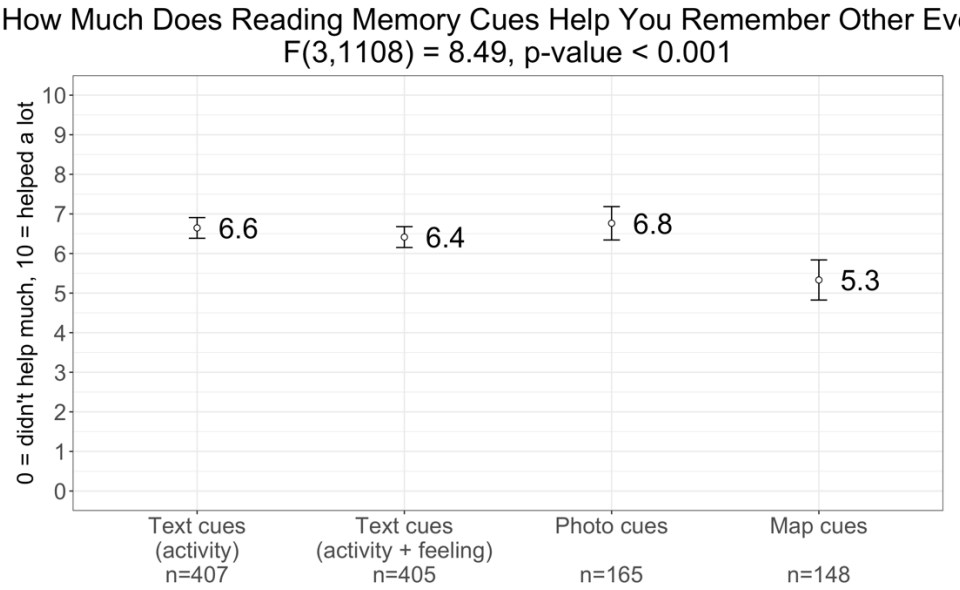


Figure 4.2 Perceived Helpfulness of Cues by Cue Type
(The F-test is based on the ANOVA model shown in Table A4.5, controlling for age, gender, education level, day of the week, self-rated memory, and memory capability.)

[13] Due to a programming error in the first half of the data collection, this question was shown to those respondents in the photo and map cue conditions. This error was fixed in the second half of the data collection, so the sample size of these two conditions are about half of the size.

Memory Capability (RQ3)

When breaking down the aforementioned four measures by the memory capability indicator—whether a respondent reported 12 or more events in the baseline time diary without any cues—we found cues had no impact on individuals who have low memory capability but had a negative impact on individuals who have higher memory capability (see Figure 4.3 below). We originally hypothesized that if the cues had a positive impact on recalling uncued events, they would have a stronger effect on individuals who have weaker memories, just as we observed for the cued events (Chapter 3). The left panel in Figure 4.3 shows that cues had no effect even on individuals who have low memory capability; the panel on the right indicates that cues had a negative impact on recalling or/and reporting uncued events for individuals who have stronger memory capability. The reasons for this are unknown.

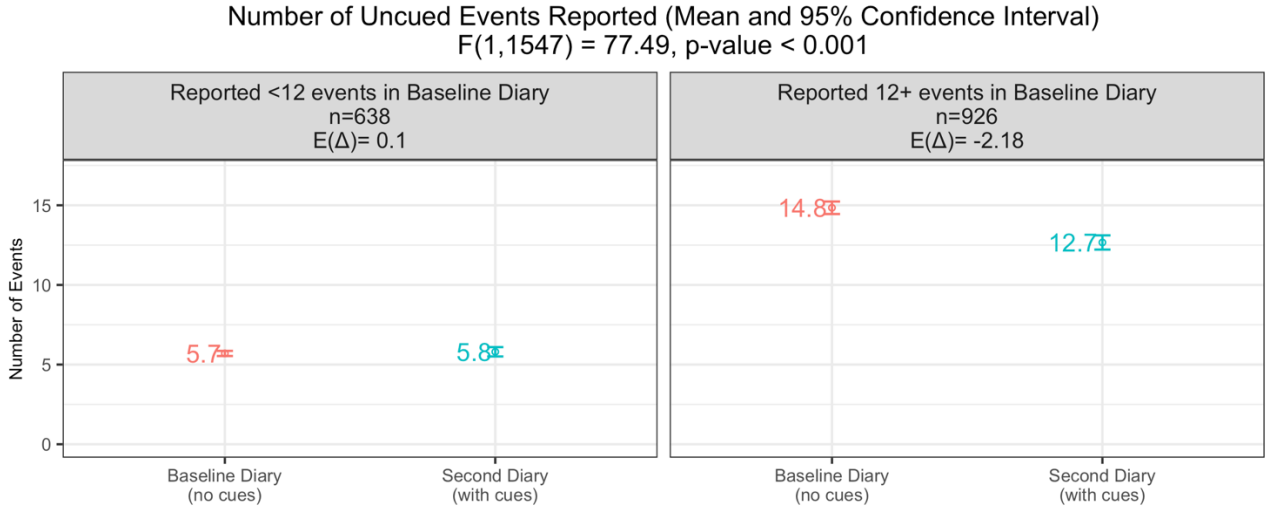


Figure 4.3 Number of Uncued Events Reported by Memory Capability
(The F-test is based on the ANOVA model shown in Table A4.1, controlling for cue type, age, gender, education level, day of the week, and self-rated memory.)

The groups of respondents gave equivalent subjective ratings when assessing how much cues helped them remember other activities (see Figure 4.4; F(1,1108) = 0.11, p = 0.746, controlling for cue type, age, gender, education level, day of the week, and self-rated memory). The two groups did not differ in their evaluations of perceived helpfulness, which indicates the negative effect is not about the embedded cues harming a respondent's memory but about the embedded cues in the instrument somehow introducing interference into the reporting process.

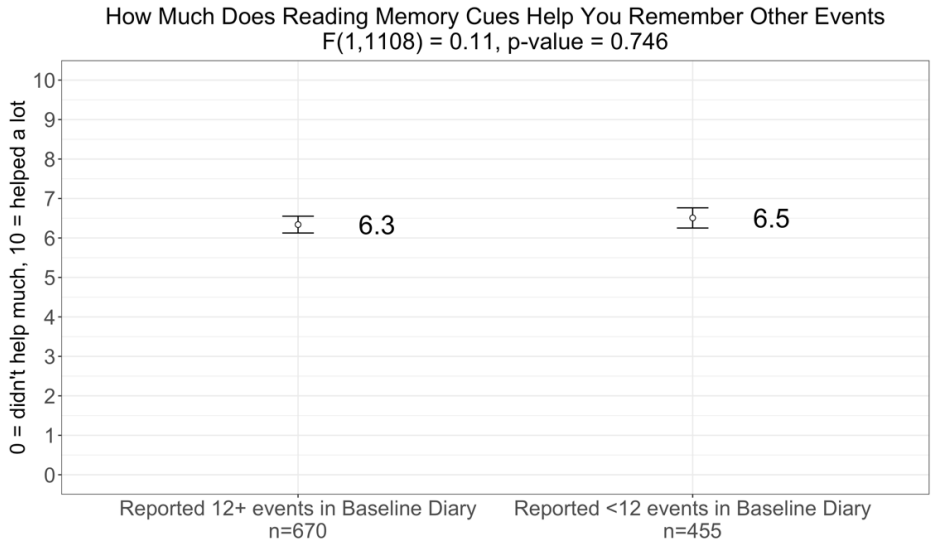


Figure 4.4 Perceived Helpfulness of Cues by Memory Capability
(The F-test is based on the ANOVA model shown in Table A4.5, controlling for cue type, age, gender, education level, day of the week, and self-rated memory.)

Reporting Effort on Uncued Events Among MB Conditions

In completing the second time diary, respondents spent 17 fewer seconds on reporting uncued events (t(1557) = -5.2, p <0.001, see Figure 4.5), which to some degree explains why the data quality of uncued events seems to be worse than that of cued events. This is aligned with the zero-sum game observation: respondents seemed to be willing to invest a fixed amount of effort into the overall time diary task, meaning that if they invested more effort in some components of the task (e.g., processing cues and carefully reporting cued events), they would invest less effort in the remaining components of the task (i.e., remembering and reporting other uncued events).

This might also be due to the cued events being more salient, leading respondents to consider them to be more important and reducing the respondents' diligence when reporting uncued events.

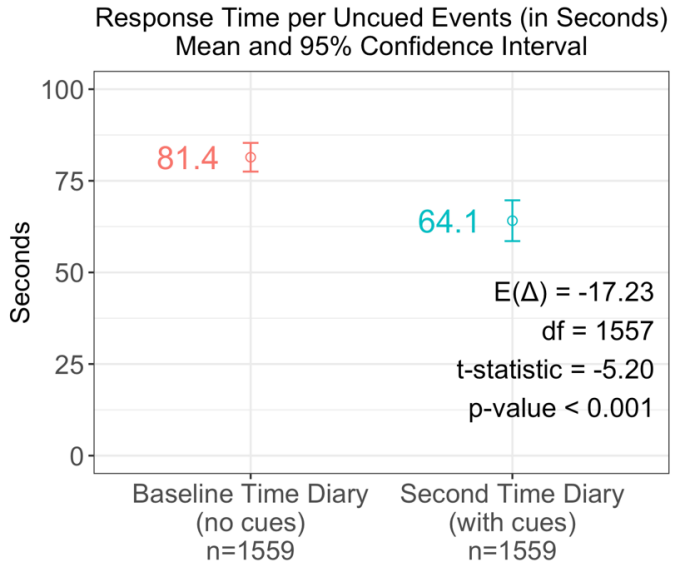


Figure 4.5 Response Time per Uncued Event (in Seconds)
(E(Δ) indicates the average difference between the baseline time diary and second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1.)

Effect on Overall Time Diary Data Quality (RQ4)

The seemingly zero-sum game phenomenon—a positive effect on cued events and a negative effect on uncued events—represents a mixed result for the entire time diary using the MB approach (see Figure 4.6 below). On the positive side, when MB respondents were cued, their time diaries were more complete than were their baseline time diaries: overall, the gaps between events were 7 minutes shorter when they were cued than when they were not ($t(1563) = -2.30$, $p = 0.022$). Yet their confidence in their reports was no worse in the second than in the first time diary. The accuracy of reported locations in the second time diary was 3% lower ($t(1563) = -2.30$, $p = 0.022$). They reported 0.8 fewer events in total ($t(1563) = -6.27$, $p < 0.001$); the average duration of reported events was 7 minutes longer ($t(1563) = 4.58$, $p < 0.001$); and the average length of the

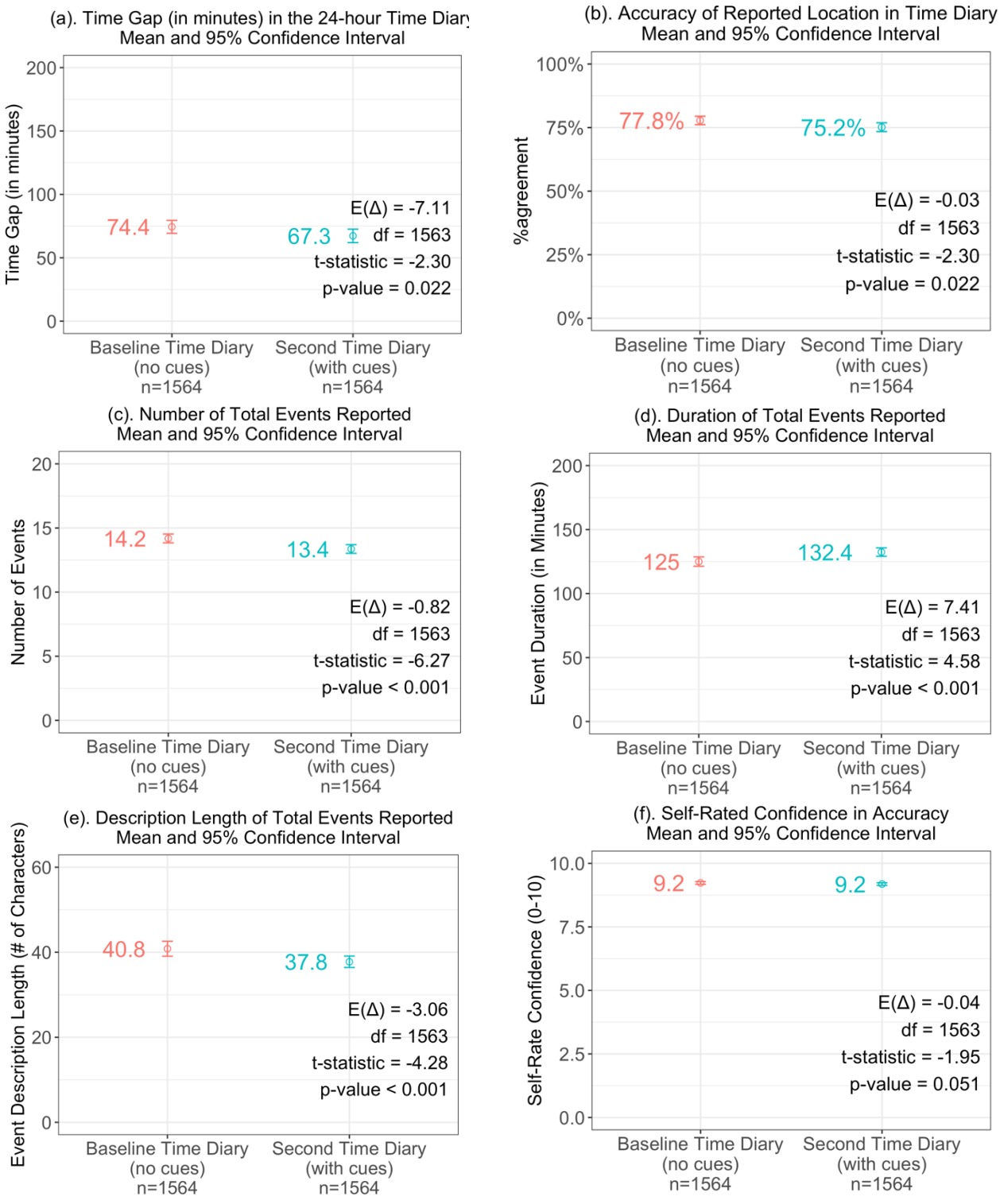


Figure 4.6 Data Quality Measures of the Entire Time Diary (MB Conditions)
(E(Δ) indicates the average difference between the baseline time diary and second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1)

event description was 3 characters shorter (t =-4.28, p <0.001). All of these differences are relatively small in magnitude, but they are statistically significant (type II error = 0.05).

The negative effect of cues on the four data quality measures were all due to uncued events. Considering that the number of uncued events was much larger than that of the cued events (i.e., the number of cued events in this study was a maximum of 4, but individuals reported on average 16 events in their entire time diaries), the positive effect on cued events was not enough to cancel out the negative effect on uncued events. There was one exception: respondents whose memory capability is lower seemed to benefit from this approach. These respondents reported 0.8 more events (9.2 events in total) in the entire time diary because of the gains in cued events and the neutral effect in uncued events. Nevertheless, the magnitude of the improvement was small, and the overall data quality of their time diaries was still substantially worse than that of the rest of the sample who reported an average of 16.2 events in the second time diary.

Overall, respondents did not perceive any more burden in completing the second time diary than they did in completing the baseline time diary (i.e., without cues). They rated the perceived work of completing each time diary a 5 (out of 10), where 10 = "a lot of work" (t(1553) = 0.5, p = 0.618; see Figure A4.2).

Control Condition and Practice Effect

To evaluate the effect of the proposed MB approach on the entire time diary (including both cued and uncued events), we included a control condition in which the respondents had no cues in either of their diaries. Our motivation for including this condition was to confirm that, in the absence of cues, data quality was stable across the two time points. However, the result turned out to be quite different. The second time diary of the control group showed an overall temporal

gap reduced by about 40 minutes (t(411) = -6.90, p <0.001; see Figure 4.7a); 5% more accuracy on reported location (t(411) = 2.26, p = 0.024; see Figure 4.7b); and 0.53 more reported events on average (t(411) = 2.04, p = 0.042; see Figure 4.7c). The granularity of the events was the same in both control time diaries, with an average event duration of 119 minutes (t(411) = -0.13, p = 0.894; see Figure 4.7d). We hypothesize that this is a practice effect: respondents carried out the exact same task twice, performing slightly better the second time.

Considering that the results of the MB conditions did not clearly improve in the second time diary, we did not further compare it to the control condition. It seems likely that completing

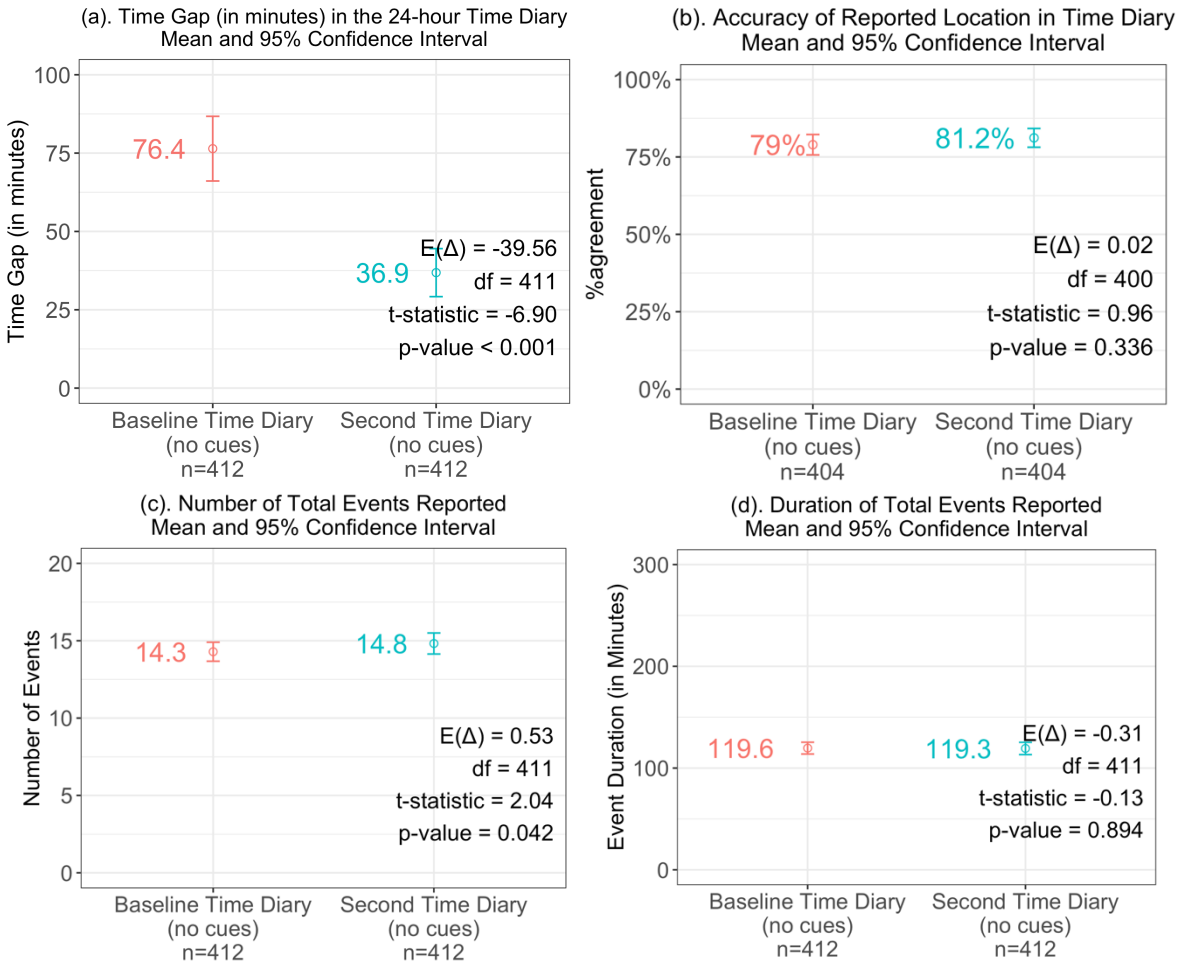


Figure 4.7 Data Quality Measures for the Entire Time Diary (Control Condition Only)
(E(Δ) indicates the average difference between the baseline time diary and the second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1

the second time diary with cues was perceived as a fundamentally different task than completing the baseline time diary without cues.

4.5 Discussion

Zero-sum game

We found the MB approach to have neutral to negative effects on the reporting of uncued events. We observed higher-quality reports of cued events but equivalent- or lower-quality reports of uncued events. This make senses when considering the respondents' perspective; it seems that they were willing to put a fixed amount of effort into completing a survey task. If they put more effort into some subtasks, then they would put less into others. The resulting net impact on the entire time diary, then, depends on the sum of the effects on cued and uncued events. For instance, respondents who have low memory capability benefited the most with respect to cued events (see Figure 3.4 in Chapter 3), but they were not affected when recalling uncued events (see Figure 4.7). Therefore, they reported slightly more events overall in the full 24-hour time diary. However, that is the only advantage we found for the MB approach; for other measures, the approach had no impact or a negative impact.

However, we saw some evidence suggesting that our findings were caused by the particular way in which this study was executed. Although the respondents who reported 12 events or more in the baseline time diaries reported substantially fewer uncued events (Figure 4.3) in their second time diaries, they reported in the debriefing that the cues "somewhat" helped them recall the uncued events (Figure 4.4). If the negative effect on uncued events was mostly caused by the design of the web interface (e.g., requiring people to answer questions about cued events first and uncued events second) and if the design is amendable in future iterations, then we might start to see an overall gain. In any event, the zero-sum game metaphor fits the data well at this point.

Future Directions

Before talking about specific recommendations for future studies, we first would like to emphasize the challenges of implementing time diaries using mobile devices. In this study, about 49% of the respondents used their smartphones to complete both time diaries; 29% used their personal computers (PCs) for both; and the remainder of the 22% switched from one to the other in the two time diaries. The data quality was much better overall for those time diaries that had been completed using PCs than those that had been completed using mobile devices (see Figure A4.3c). Looking at the quality of the data generated by respondents who did not switch devices between time diaries, respondents reported 2.5 more events when using their PCs when compared to their mobile devices ($t(1054) = 7.6$, $p < 0.001$); moreover, the accuracy of reported locations increased by 6.2% ($t(1315) = 4.67$, $p < 0.001$). More data quality measures and figures can be found in Figure A4.3. Providing data on a mobile device is certainly more challenging than it is on a PC when the survey task is lengthy and complex, as is a time diary. The proposed MB approach adds another layer of complexity to this.

In considering future time diary studies that may use the MB idea, we provide the following recommendations to improve the process based on what we have learned:

- *Make the cue information salient and easy to access*. In the current design, cues (text, photo, or map) are nested within the cued events. Respondents need to click several times to access the cue information after they had filled out the cued events. This is far from ideal. We envision a design in which the cue (text, photo, or map) "floats" on top of the time diary. This would make the cue easy to open or hide directly on the surface. We have not figured out how one might do this on a smartphone, but for respondents using a bigger screen, floating interface objects could reduce clutter and simplify the

task. We anticipate that a mobile application may better implement this idea, which is worthwhile investigating.

- *Prompt respondents to be aware of temporal gaps*. Without cues, respondents are more likely to remember and report events in a chronological order, so it is rare for them to accidentally forget to report some events they can remember. However, when the provided cue information is sparse, i.e., distributed throughout the day, respondents may not reconstruct their days in a chronological order, thus creating more temporal gaps between events. Highlighting and prompting the temporal gaps between events should increase respondents' awareness of missing events.
- *Pick whatever cue works best for the respondents.* In this experiment, we saw some advantage to photo cues but not a lot. For the best results, future studies might consider letting respondents choose whichever cue works best for them or record any piece of information that will help them best encode what is happening at the moment. A previous study showed that giving individuals a choice about how they respond to a survey could improve performance (Conrad et al., 2017).
- *MMS is not the best way to collect photo cues.* We decided to use text messaging because of its ubiquity across populations but collecting photos via MMS is cumbersome. The transmission speed can be frustrating to users if the cellular signal is weak, which is a problem in "urban canyons" or very rural areas. We were unable to solve this problem given our resources. However, it would be ideal if future studies were able to collect photo cues via Wi-Fi either using texting or a mobile application. This would represent a major improvement over the current design.

4.6 Limitations

Validating self-reported events against actual events is extremely hard in a naturalistic survey setting such as this one, given that we had no records of participants' personal events. In this study, we used up to two random GPS coordinates as a benchmark of a single event. However, we still required some subjective input: we asked respondents to label what the locations meant to them after completing their time diaries. Two random GPS benchmarks are better than no benchmarks but still might not represent accuracy over an entire day. If it was possible to continuously track respondents' GPS coordinates for a whole day (e.g., by using a mobile app), then we would have location benchmarks for all the reported.

In this study, we used a convenience sample rather than a probability-based sample due to financial constraints. However, because we balanced the key demographics of gender, age, education level, day of the week, self-perceived memory, and memory capability. We also used repeated measures design and thus the results of the study are reasonably generalizable to the broader population.. Nevertheless, we had very little control over respondents' demographics, and we do not know if randomly recruited participants would produce comparable results.

References

Bureau of Labor Statistics. (2016). American Time Use Survey questionnaire – 2011-2015. Retrieved from http://www.bls.gov/tus/tuquestionnaire.pdf

Conrad, F. G., Schober, M. F., Antoun, C., Yan, H. Y., Hupp, A. L., Johnston, M., ... & Zhang, C. (2017). Respondent mode choice in a smartphone survey. *Public Opinion Quarterly*, *81*(S1), 307-337.

Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., & Toepoel, V. (2019). Where You at? Using GPS Locations in an Electronic Time Use Diary Study to Derive Functional Locations. Social Science Computer Review. https://doi.org/10.1177/0894439319877872

Fernee, H., Sonck, N., & Scherpenzeel, A. (2013). Data collection with smartphones: experiences in a time use survey.

Fisher, K., & Gershuny, J. (2013). *Time use and time diary research*. Oxford University Press.

Chatzitheochari, S., Fisher, K., Gilbert, E., Calderwood, L., Huskinson, A., Cleary, A., & Gershuny, J. (2015). *Measuring young people's time use in the UK millennium cohort study: A mixed-mode time diary approach*. CLS Working Paper 2015/5. London: Centre for Longitudinal Studies, University of London.

Hamermesh, D. S., & Pfann, G. A. (2005). Time-use Data in Economics. European Economic Review, 49(1), 1-7.

Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., ... & Wood, K. (2006, September). SenseCam: A retrospective memory aid. In *International Conference on Ubiquitous Computing* (pp. 177-193). Springer, Berlin, Heidelberg.

Kelly, P., Thomas, E., Doherty, A., Harms, T., Burke, Ó., Gershuny, J., & Foster, C. (2015). Developing a method to test the validity of 24 hour time use diaries using wearable cameras: a feasibility pilot. *PLoS One*, *10*(12), e0142198.

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., & Liss, S. (2011). Summary of travel trends: 2009 national household travel survey (no. fhwa-pl-ll-022). Retrieved from http://nhts.ornl.gov/2009/pub/stt.pdf.

Statistics Bureau, Japan (2016). 2016 Survey on Time Use and Leisure Activities. Retrieved from http://www.stat.go.jp/english/data/shakai/2016/pdf/qua.pdf

Statistics Canada (2016). General Social Survey – Time Use (GSS). Retreived from http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4503#a2

Stone, A. A., & Mackie, C. (Eds.). (2014). *Subjective well-being: Measuring happiness, suffering, and other dimensions of experience*. National Academies Press.

United Nations (2018). *Gender Statistics*. Retrieved from https://unstats.un.org/unsd/gender/timeuse/

Appendices

Appendix 4.1 Additional Tables

Table A4.1 ANOVA Model of Difference in Number of Reported Uncued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	128.712	3	1.727	0.160
Age	208.327	3	2.795*	0.039
Gender	37.609	1	1.514	0.219
Highest Education Degree	99.856	2	2.009	0.134
Day of the Week	39.226	1	1.579	0.209
Self-rated Memory	1.692	2	0.034	0.967
Memory Capability	1925.599	1	77.491***	<0.001
Device	208.589	3	2.798	0.039
Residual	38441.749	1547	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.2 ANOVA Model of Difference in Average Event Duration of Uncued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	32519.534	3	2.132	0.094
Age	4287.524	3	0.281	0.839
Gender	198.158	1	0.039	0.844
Highest Education Degree	57972.760	2	5.700*	0.003
Day of the Week	39.246	1	0.008	0.930
Self-rated Memory	6558.297	2	0.645	0.525
Memory Capability	16366.655	1	3.218	0.073
Device	64546.716	3	4.231**	0.005
Residual	7841465.671	1542	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.3 ANOVA Model of Difference in Average Event Description Length of Uncued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	9120.213	3	4.856**	0.002
Age	3833.744	3	2.041	0.106
Gender	1767.273	1	2.823	0.093
Highest Education Degree	4310.322	2	3.442*	0.032
Day of the Week	18.624	1	0.030	0.863
Self-rated Memory	597.481	2	0.477	0.621
Memory Capability	76.920	1	0.123	0.726
Device	15206.214	3	8.096***	<0.001
Residual	965425.697	1542	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.4 ANOVA Model of Difference in Average Self-Rated Confidence of Uncued Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	5.729	3	1.837	0.138
Age	5.599	3	1.795	0.146
Gender	1.296	1	1.247	0.264
Highest Education Degree	4.945	2	2.379	0.093
Day of the Week	5.530	1	5.320*	0.021
Self-rated Memory	1.275	2	0.613	0.542
Memory Capability	1.071	1	1.030	0.310
Device	2.412	3	0.774	0.509
Residual	1602.846	1542	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.5 ANOVA Model of How Cues Helped Respondents Remember Other Events

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	193.153	3	8.486***	<0.001
Age	50.322	3	2.211	0.085
Gender	2.909	1	0.383	0.536
Highest Education Degree	52.510	2	3.460*	0.032
Day of the Week	4.468	1	0.589	0.443
Self-rated Memory	2.008	2	0.132	0.876
Memory Capability	0.798	1	0.105	0.746
Device	86.603	3	3.805*	0.010
Residual	8406.857	1108	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.6 ANOVA Model of Difference in Number of Reported Total Events Between the Two Diaries

	Sum of Squares	DF	F value	Pr(>F)
Cue Type	131.169	3	1.773	0.150
Age	241.588	3	3.266*	0.021
Gender	46.334	1	1.879	0.171
Highest Education Degree	89.850	2	1.822	0.162
Day of the Week	27.326	1	1.108	0.293
Self-rated Memory	1.685	2	0.034	0.966
Memory Capability	2931.300	1	118.887***	<0.001
Device	172.669	3	2.334	0.072
Residual	38143.064	1547	NA	NA

Note: Dependent Variable = Second Time Diary – Baseline Time Diary
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table A4.7 Linear Regression Results Assessing the Effect of Cue Type and Other Blocking Variables on the Number of Uncued Events

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	-2.160***	0.596	-3.622	<0.001
Cue Type				
Text Cue (Activity only)	--	--	--	--
Text Cue (Activity + Mood)	-0.609	0.347	-1.755	0.079
Photo Cue	0.021	0.356	0.058	0.954
Map Cue	-0.503	0.360	-1.398	0.162
Age				
18–24 years old	--	--	--	--
25–34 years old	0.488	0.355	1.373	0.170
35–44 years old	1.099	0.395	2.785	0.005
45+ years old	0.337	0.464	0.725	0.469
Gender				
Male	--	--	--	--
Female	-0.342	0.278	-1.230	0.219
Highest Education Degree				
Graduate degree	--	--	--	--
College degree	-0.615	0.355	-1.731	0.084
No college degree	-0.148	0.355	-0.416	0.678
Day of the Week				
Weekday	--	--	--	--
Weekend	0.322	0.256	1.256	0.209
Self-rated Memory				
Excellent	--	--	--	--
Very good	-0.031	0.360	-0.088	0.930
Good/Fair/Poor	-0.077	0.303	-0.254	0.800
Memory Capability				
12+ events in Baseline	--	--	--	--
<12 events in Baseline	2.334***	0.265	8.803	<0.001
Device				
Mobile for both	--	--	--	--
PC for both	0.407	0.305	1.334	0.182
Mobile to PC	0.489	0.370	1.322	0.186
PC to Mobile	-1.094	0.563	-1.945	0.052

Note: Dependent Variable = Second Time Diary – Baseline Time Diary, N = 1559, Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Appendix 4.2 Additional Figures

Figure A4.1 Data Quality Measures of Uncued Events by Cue Type
(ANOVA tests the difference between two time diaries across four conditions in a model that controls for age, gender, education, day of the week, self-rated memory, and memory capability.)

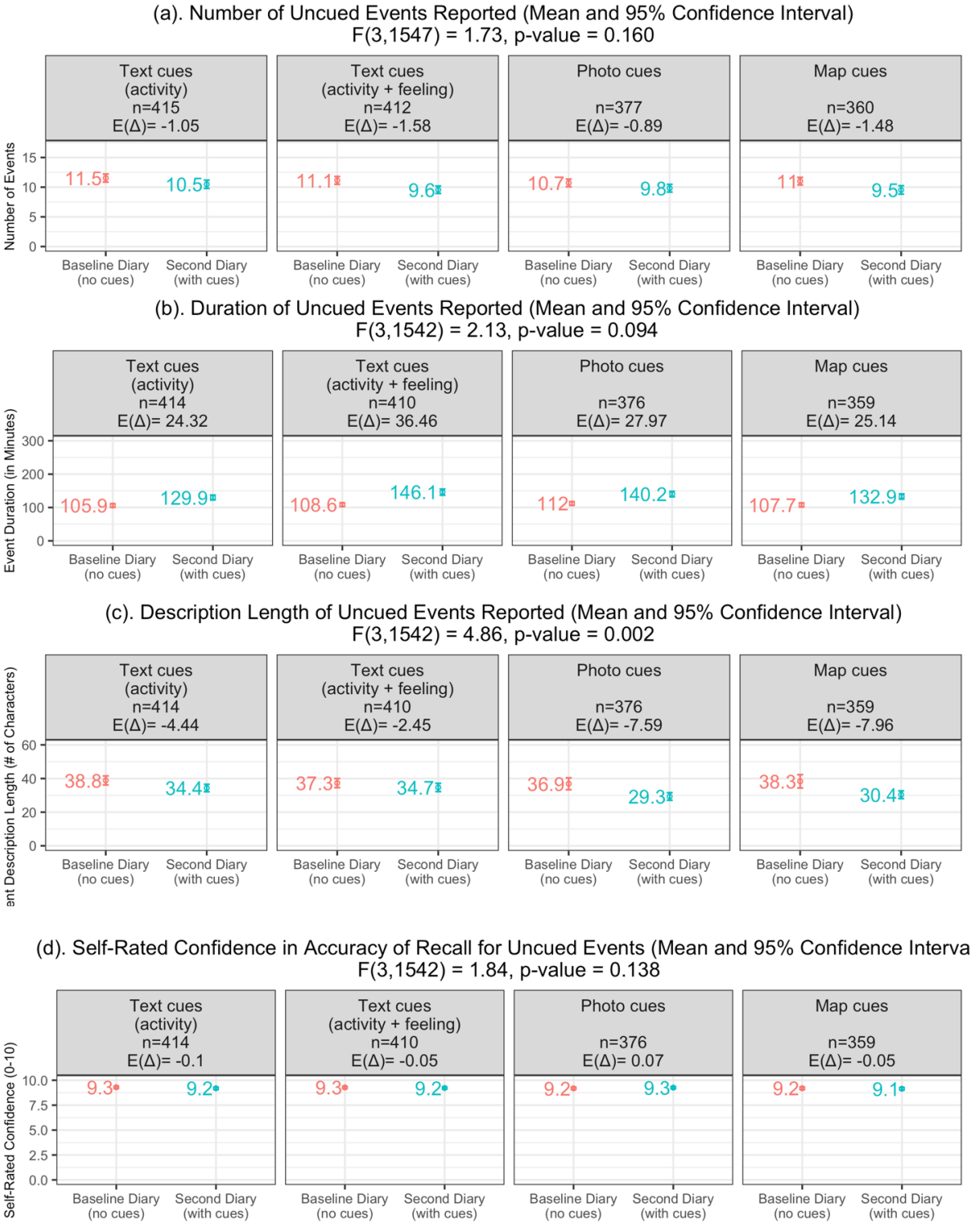


Figure A4.2 Perceived Work in Completing Entire Time Diary
(E(Δ) indicates the average difference between the baseline time diary and the second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1)

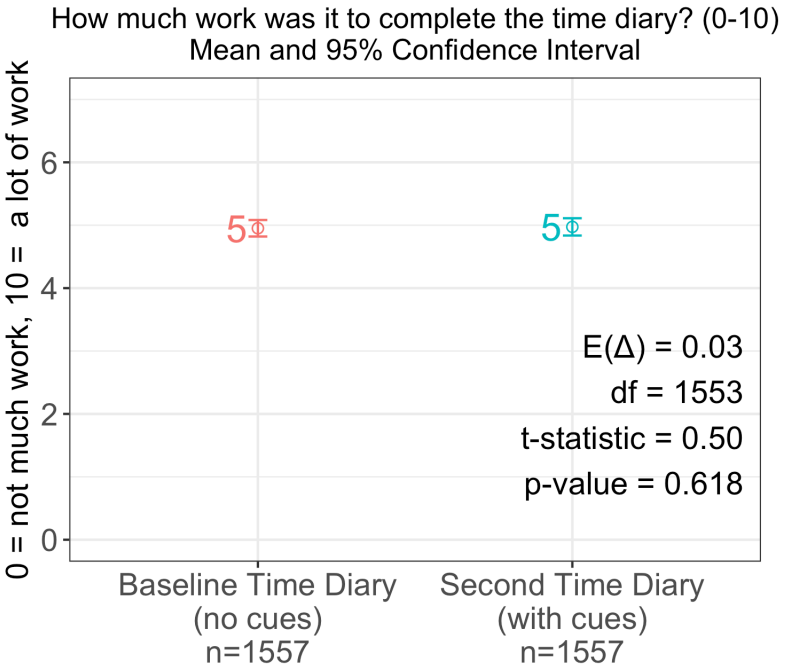
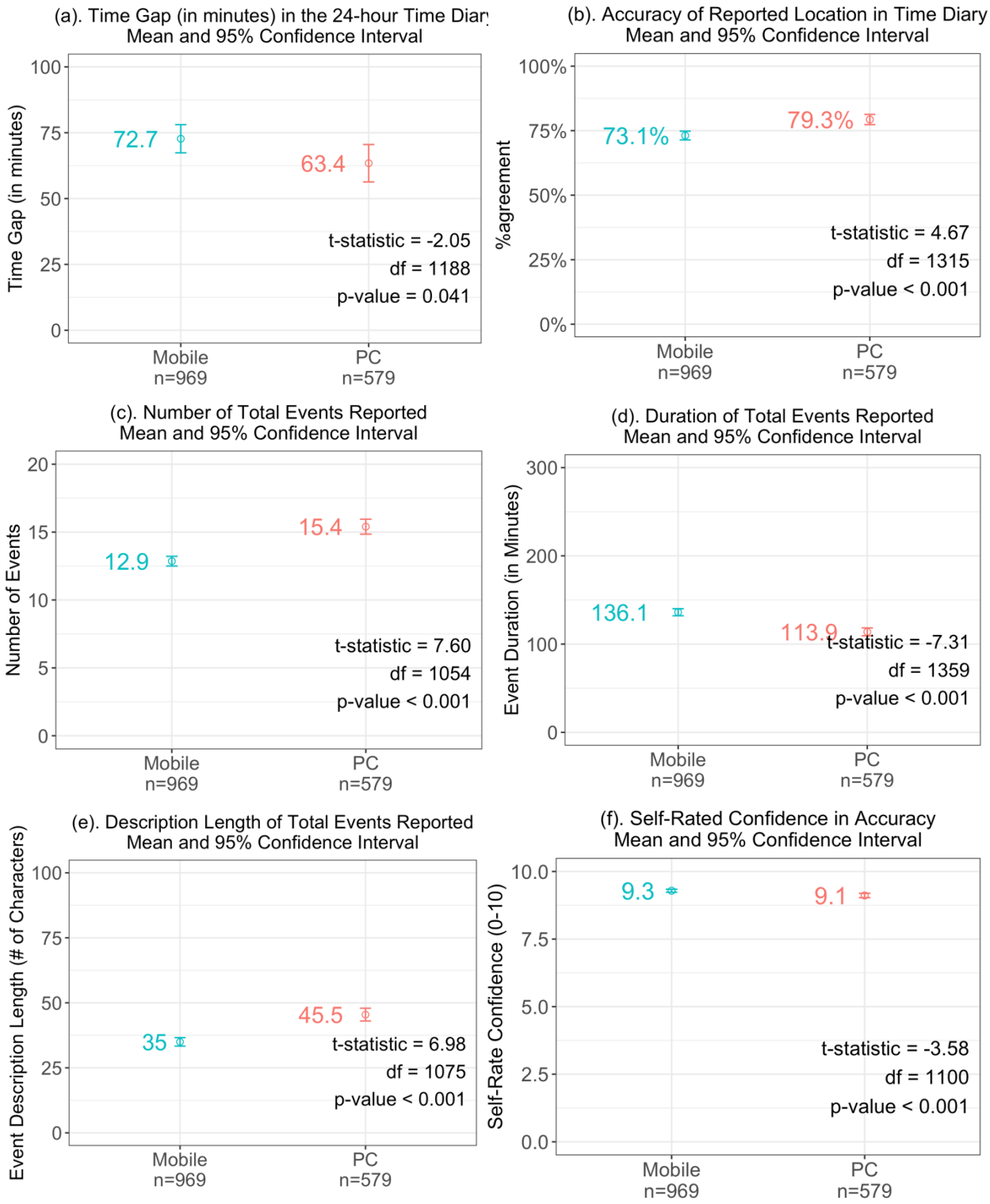


Figure A4.3 Data Quality Measures for the Entire Time Diary
(Restricted to respondents who did not switch devices between the two time diaries, E(Δ) indicates the average difference between the baseline time diary and the second time diary per respondent; the t-statistics are based on a one-sample t-test of E(Δ); degree of freedom = n-1)



Chapter 5 Conclusions

Individuals' memories are not always reliable when it comes to answering survey questions about past behaviors. There is a general consensus that memory error as well as the cognitive burden of recall should be reduced during survey data collection. The existing data collection methods vary in the degree to which they rely on recall. One extreme is the traditional retrospective approach, which relies solely on a respondent's memory. In this study, this approach is represented by the baseline time diary, in which we asked respondents to recall what had taken place in the past 24 hours without using any cues. The other extreme is the elimination of recall during data collection. Examples of this latter method are intensive longitudinal approaches such as the experience sampling method (ESM) and ecological momentary assessment (EMA), which ask respondents to answer all questions of interest *in situ* many times during a single day. There is another growing research stream on passive data collection using sensors on wearable devices, which requires respondents' permission to access their logged data (e.g., Doherty et al., 2017).

In this dissertation, we proposed and executed a new data collection method—the memory bookmarking (MB) approach—that falls in between the two aforementioned extremes. Put simply, the MB approach turns free recall into cued recall by using *in situ* information from respondents. In this study, we first collected very minimal *in situ* information via texting, which does not require recall or much input from the respondents (two clicks for recording a GPS location, taking one photo, or replying to one text message). We later used this *in situ* information as recall cues to prompt for more details in a follow-up survey.

We conducted a pilot study using the MB approach with text cues and a follow-up time diary to demonstrate the feasibility of the two-step data collection framework (Chapter 2). We obtained some preliminary findings with regards to both experimental design and system design that we were able to use in designing the main study. In the main study, we evaluated two kinds of text cues, as well as photo cues and map cues, for promoting recall in online data collection.

In Chapter 3, we found that such recall cues helped individuals remember and report more cued events than they did when engaged in free recall without cues. Regardless of cue type, events reported using cues were more fine-grained than events reported without cues. Map cues require the least effort to create but might provide the least information for recall compared to the other types of cues (two types of text cues and photo cues). Subjectively, the respondents were more confident about their recall when photo cues were present. Overall, the respondents said it took very minimal work to reply to these cues, which affirms the feasibility of using such an approach in future studies. This study also provides evidence that the MB approach works better for individuals whose memory ability is relatively low. In our future work, we would like to give respondents the opportunity to choose the type of cues that would benefit them the most and investigate the effect of the chosen cues.

In Chapter 4, we further evaluated whether cues could help respondents recall uncued events (adjacent but not cued events) in a 24-hour time diary. We found a neutral to negative effect for these events. We saw higher-quality reports of cued events (Chapter 3) but equivalent- or lower-quality reports of uncued events, making for a seemingly zero-sum game. This makes sense from the respondents' perspective; it seems they were only be willing to put a fixed amount of effort into completing a survey task. If they put more effort into some subtasks, they put less into others. The resulting net impact on the entire time diary depended on the sum of the effects on

cued and uncued events, and it varies by respondents. For instance, respondents who have low memory capability benefited the most with respect to cued events; however, they were not affected when recalling uncued events. Therefore, they reported slightly more events in the full 24-hour time diary using the MB approach. Other respondents who showed a negative impact from the cues might have done so because of the particular way that the web instrument was designed. Subjectively, respondents thought the cues were somewhat helpful to remember other events although we did not find quality gains in their actual reports. For time diary application, future studies should look into improvements that could be made to the MB approach (see detailed recommendations in the *Discussion* section in Chapter 4) before definitively concluding that the MB approach has no impact or a negative impact on individuals who have relatively higher memory capability for recalling their personal activities in a time diary.

References

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M H., ... & Preece, S. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one*, *12*(2), e0169649.