What Are You Doing Now? Activity Level Responses and Recall Failures in the American Time Use Survey

Tarek Al Baghal
Institute of Social and Economic Research (ISER)
University of Essex
Colchester, UK

Robert F. Belli
Survey Research and Methodology
University of Nebraska-Lincoln
Lincoln, Nebraska

A. Lynn Phillips
Survey Research and Methodology
University of Nebraska-Lincoln
Lincoln, Nebraska

Nicholas Ruther
Survey Research and Methodology
University of Nebraska-Lincoln
Lincoln, Nebraska

Author contact: talbag@essex.ac.uk

ACKNOWLEDGEMENT: *This material is based upon work supported by the National Science Foundation under Grant No. 1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.*

Word Count: 5999

Activity Level Responses and Recall Failures in the American Time Use Survey

Abstract

Questions about people's pasts are common in many surveys, but memories are error prone. The current research focuses on recall failures in the American Time Use Survey (ATUS). The ATUS most commonly encourages respondents to report all of their activities of the previous day in a forward chronological fashion, from the beginning to the end of the day. Even with a short reference period, the ATUS is prone to recall errors. We explore these errors taking into account the response process, respondent, and interviewer as possible contributors to a recall failure. Importantly, we posit that the chronological recall of events leads to earlier activities affecting recall of the current activity. Events are more easily recalled when they are more distinct (less frequent) or additional contextual information about the event is available. While research has focused on these characteristics of the target event, the previous event recalled may also provide distinctiveness and context. Results suggest that periods following a more frequent activity are likely to be followed by a failure, although this is modulated by the duration of the event. The presence of others and places of the event also have significant effects. The elapsed time since the event is also important, with a higher chance of recall failure for more distant activities. Although results highlight the importance of the response-level in understanding outcomes, respondent characteristics still matter, as those with apparently lower cognitive ability are more likely to have a failure. Interviewers also contribute to the variance of recall failures, with interview experience not having an apparent effect, while interviewers who make other types of errors, surprisingly, show lower likelihoods of recall failure. The results shed light on the relationship between memory and survey errors, and suggest implications for future survey design.

1. Introduction

Numerous studies and surveys are interested in respondents' pasts, including how people use their time. Three main methods are used to collect time-use data: experience sampling (e.g., Hektner et al 2007), stylized questions, and time-use diaries. Whereas experience sampling techniques ask about events in real-time, both stylized questions and time-use diaries require autobiographical recall of activities and their durations in a given reference period (Juster 1985). When dealing with autobiographical memories, errors often occur (Thompson et al. 1996; Tourangeau et al. 2000), which is also the case in those reports that produce time-use data (Sturgis 2004; Fricker 2007; Freedman et al. 2012; Phillips et al. 2013).

Time-use diaries have been established as a reliable source of data (Michelson 2005), roughly comparable to experience sampling methods (and less expensive), and more valid than stylized questions (Bolger et al. 2003). Other advantages of time-use diaries include the ability to collect data about the context of events: the activities that followed or preceded each event, who was present with the respondent and the location for each event (Harvey and Royal 2000). One such time-diary survey is the American Time Use Survey (ATUS). Since 2003, the ATUS has been a valuable source of information for academics, lawyers, governmental agencies, and the press. Conducted by the US Census Bureau for the Bureau of Labor Statistics, the ATUS is designed to be a high-quality, probability-based survey that is representative of the United States population.

In part, the data quality advantages of time diaries like the ATUS are due to the shortness of the reference period. Forgetting is minimized as ATUS interviewers only ask about activities that had occurred the previous day (Abraham et al. 2006), although some researchers argue that less memorable events that happen earlier in the day will not be remembered (Hektner et al.

2007). Data quality in the time diaries is also promoted by the temporal and thematic relatedness of the activities that are reported. In a diary questionnaire such as the ATUS, recall of events typically happens in a forward sequential manner, in which every earlier activity of the prior day is reported immediately before the later, and adjacent, activity. Temporal linkages among adjacent events, along with thematic relationships, are known to structure autobiographical memory (Barsalou 1988; Conway 1996) and improve the data quality generated in calendar interviews (Belli 1998; Belli et al. 2009). In the ATUS, the recall of one activity serves as both a temporal and thematic cue in remembering the next (Stafford 2009).

The task for respondents answering time diary questionnaires is to report on the actual and specific activities that happened yesterday, and accordingly, they are asked to rely on their episodic memory to recall these activities. Episodic memory depends on events that are distinctive from other events, which permits them to be located reliably in memory (Burton and Blair 1991; Menon 1993). One threat to accurate recall is an under-reliance on episodic memory complemented by an overreliance on generic memory, which involves remembering what typically occurs (Linton 1982; Means and Loftus 1991). For example, respondents who often engage in daily routines may find it difficult to disentangle what exactly happened yesterday from what typically happens at certain times of the day. Accordingly, generic memory may be relied on when events lack distinctiveness, which often results in decrements to data quality when the task requires the reporting of specific episodes.

In this research, we take a different tack in examining the role of distinctiveness on data quality. Instead of focusing on the impact of distinctiveness in reporting an activity, we examine the impact of distinctiveness in reporting the next activity. We hypothesize that the amount of distinctive information of what is remembered from each earlier activity will determine its cue-

effectiveness for remembering the immediately following adjacent activity. Distinctive information may be contained in the content of what is remembered, or may involve the extent to which the temporal location of an activity's occurrence is distinctive. Our hypothesis is based on the notion that linkages among temporally adjacent activities are more robust when there is a greater level of distinctive information available in the earlier activity to cue an episodic memory of what happened next.

To examine this hypothesis, we focus on characteristics of activities that will be associated with the amount of distinctive content during remembering, and determine whether these characteristics are predictive of the ability to report on the following adjacent activity. One of these characteristics is the frequency of occurrence of an activity. Activities that occur frequently are those that are more likely to be routine parts of everyday life (Linton 1981). Their memory will be stored in terms of what typically happens rather than in terms of what specifically happened, and hence are likely to be lacking in both content and temporal distinctiveness.

A second characteristic is duration. In autobiographical memory, recall of events is facilitated at the transition points between long extended events (Pillemer et al. 1988, Robinson 1986). Similarly in time diaries, long duration activities may lead the transition to the next activity to be more memorable in comparison to activities of short durations. Simply by virtue of their longer duration, such activities may contain more potentially recoverable distinctive details than shorter events (Brown 1997). Finally, some long duration activities will no doubt exceed the usual lengths of activities of that type, making them more distinctive. Additional findings also show that the presence of others and where an event occurred can aid in cuing recall (Brewer 1988, Wagenaar 1986) by increasing their distinctiveness (e.g. Brown 1995, 1997).

The ATUS dataset provides a unique opportunity to test the cuing potential of various types of activities as well as other factors such as the presence of other individuals and the location of activities on respondent recall. Findings from such analysis are important to not only time-diary survey research, but more generally for understanding how the nature of memory can affect survey data quality.

2. Data and Methods

The data come from the 2010 American Time Use Survey (ATUS). The ATUS sample is drawn from those who are selected for the Current Population Survey (CPS) in a three-stage stratified sample.[1] Samples of each week are split to be conducted equally on weekend and weekday days. Households are sent an advance mailer and contacted and interviewed by telephone. Households without a telephone are sent an advance mailer with a call-in telephone number and a $40 incentive. Household members 15 years old and older are eligible for selection to complete the interview, conducted in either English or Spanish. The ATUS interview has several components, with the time diary of central analytic focus in the current work. The time diary portion uses conversational interviewing rather than using scripted questions.

Respondents are asked to report all activities and timing of these (either by giving start and stop times or duration of activity) beginning at 4 A.M. of the previous day, going forward through the day until 4 A.M. of the current day. Respondents can provide as few or as many activities as they can recall in the 24-hour period, also reporting the time and details of the event such as presence of others and place of occurrence. Although some activities are more likely to have taken place with another person present (e.g. telephone calls, caring for others), no activity

[1] For complete information, see "American Time Use Survey User's Guide" http://www.bls.gov/tus/atususersguide.pdf

occurred with either only the respondent or with others present in every instance. A number of predefined codes are used to capture where the activity occurred, including at home, in transit, at work, and other types of places.

In post-processing of the survey, responses to the time diary are coded into three tiers of activities, each with higher levels of specificity, with "first-tier" being the broadest groupings of activity types. The ATUS has 18 "first-tier" codes to which all activities can be assigned. Also in post-processing, the ATUS codes six types of errors at the activity level, which indicate different coding problems: failure to record travel (i.e. consecutive events occurring at different locations), refusals, and recall failure. These all occur in relatively small numbers, but recall failure is most clearly due to the respondent. The percentage of events that are coded for each error and respondents making at least one of each error is presented in Table 1.

Table 1. Error Rates by Activities Reported and Respondent

	Insufficient Detail	Missing Travel	Record simultaneous codes wrong	Refusal	Unable to code at first-tier	Recall Failure
% of Entries	0.78	0.18	0.14	0.02	0.03	0.26
% of Respondents	11.05	2.80	2.05	0.49	0.51	4.63

Errors other than recall errors are likely not related to memory processes or they may be multi-determined, with the respondent, interviewer error, or both responsible for the appearance of the problem. Further, these recall errors are coded distinctly from the remainder of the errors, suggesting that these more exclusively reflect recall errors while the others measure confound more than one problem. Specifically, interviewers code recall failures immediately, and only if the respondent says directly they cannot remember. Any other report is recorded verbatim, and sent to coders, who then assign a code. These coders determine whether a verbatim answer has "insufficient detail" or some other type of error.

Included in the ATUS data set is a measure of an interviewer's appraisal of the interview quality (dichotomous). There are 81 cases (0.61%) flagged by interviewer as not being good quality. There are 13179 respondents remaining in the 2010 ATUS data, with a 56.9% response rate (AAPOR RR 2). Data was collected over the course over the entire calendar year, with approximately equal numbers of interviews being conducted in each month, with a low of 7.2% of interviews being conducted in December (n=952) and a high of 9.8% of the interviews occurring in January (n = 1300). There are 69 interviewers, with widely varying number of interviews completed, with a mean of 196.9 completes (s.d. = 206.6) and range of 1 to 780 completed interviews.

3. Results

The composition of the 2010 ATUS sample is presented in Table 2.The sample is fairly representative of the American population, with women being the one demographic group somewhat overrepresented (56%). The sample is also older than the overall US population due to the ATUS including only those 15 years and older. Weekly income is capped by the BLS in its collection of the ATUS to be $2884.61, and all reported incomes higher than this are recorded as this value. The final row of Table 2 shows that about 50% of respondents reported on a weekend day, consistent with the ATUS sampling design. In addition, a measure of response propensity to the ATUS survey request is calculated following methods similar to Fricker (2007), which uses information from prior response on the CPS, allowing for information about ATUS nonrespondents (full details in Appendix A). This measure is calculated as initially reluctant respondents are found to be less thoughtful in responding (Fricker 2007, Olson 2006). Not surprisingly, on average respondents had a reasonably high propensity to respond.

Finally, two interviewer characteristics are included; the number of completes an interviewer has, used as an indicator of interviewer experience (at least with the ATUS in 2010) and the coded errors (other than recall failure) for an interviewer. More experience has been shown to have some potentially negative impacts on survey outcomes. These include faster paced surveys (Olson and Peytchev 2006) and more acquiescent responses (Olson and Bilgen 2011). Coded errors besides recall failure are used as a possible indicator of interviewer capability. Table 2 shows that interviewers in general completed a number of surveys, with large variation, and about 1 in 3 interviews conducted by the interviewer having one error (not recall failures).

. Table 2. Respondent Sample and Interviewer Characteristics

Variable	Mean/Proportion of Sample	S.E.
Respondent		
Age	46.837	0.154
Weekly Income	467.561	5.635
Female	0.561	0.004
White	0.662	0.004
Hispanic	0.131	0.003
Black	0.149	0.003
Employed	0.606	0.004
Out of Labor Force	0.326	0.004
Unemployed	0.068	0.002
< High School	0.161	0.003
High School	0.437	0.004
Undergraduate	0.288	0.004
Graduate	0.115	0.003
Partner	0.517	0.004
Weekend	0.503	0.004
Response Propensity	0.691	0.001
Interviewer		
Completes	196.94	24.87
Error Per Interview	0.348	0.055

Of interest is failure to recall a period of time in the past day, leaving a gap in the diary. There are 611 (4.63%) respondents that were unable to recall activities for at least one period of time during the previous day.[2] Although a small percent had at least one such failure, several points are pertinent. First, respondents were asked only to recall the previous day; memory failures of this type should be minimal. Nevertheless, a non-trivial number of respondents had at least one failure. Second, this memory failure is the only clearly observable *recall* error. Other errors in recall likely exist that are not easy identified, and thus the number of errors identified in this way is a conservative estimate of the total. Third, these failures can still be informative to understanding recall processes and its relation to survey design, in particular autobiographical memory and diary surveys.

One commonly used strategy for analyzing these failures is to examine the relationship between respondent characteristics such as those in Table 2 and whether the respondent made an failure or not. However, a growing literature suggests that characteristics surrounding the responses themselves are important in understanding important outcomes (e.g. Yan and Tourangeau 2008, Couper and Kreuter 2013). In addition, earlier responses can make information more accessible for later questions (Sudman et al. 1996). For these reasons, a potentially more fruitful method of analysis is a multilevel model capturing both important response- and respondent-level characteristics.

For recall failure, of the 611 respondents that had at least one recall failure, 556 (91%) had only one, 52 (8.51%) had two, and 3 (0.49%) had three, with none more than three. This variation means that there are 669 (0.26%) unique entries coded as recall failures. Table 3 presents the recall error rates for respondents and interviewers. Respondents infrequently had

[2] There was not significant differences detected in the number of respondents making a memory error by month (χ^2_{11} =18.95, n.s.). As such, it will not be considered in further analyses.

recall failures, with each respondent making on average 0.051 errors per interview, and 0.003 errors for each activity they reported. Interviewers average slightly less than 10 recall failures across all of their interviews, or 0.042 recall failures per interview. The differences between errors per interview for respondents and interviewers are due to interviewers with more completed interviews having more total errors, lowering the overall mean for interviewers.

Table 3. Recall Error Totals and Rates for Respondents and Interviewers

Recall Failure by Respondent (n=13179)	Mean	S.E.	Minimum	Maximum
Total Failures	0.051	0.002	0	3
Failure Rate	0.003	0.0001	0	0.200
Recall Error by Interviewer (n=69)	Mean	S.E.	Minimum	Maximum
Total Failures	9.696	1.695	0	57
Failure Rate	0.042	0.007	0	0.333

Also at the response-level, the timing of the event can affect recall. Although the likelihood of memory decay is decreased given the short reference period, periods at the beginning of the day are still more distant than those at the end of the day. Further, since respondents are surveyed over different times of the day, those surveyed later in the day will have a recall period more distant in memory than those surveyed earlier in the day. If prior activities do serve as cues for following activities, then better, more distinctive cues should increase the ability to recall activities while less distinctive cues should increase the chance that there is a recall failure.

The ATUS allows examination of this cuing possibility as the structure promotes a forward chronological recall of individual activities. Including both activities and error codes, there are a combined 256,105 unique entries for the set of respondents. The number of activities

across respondents varied, with a mean of 19.4 (s.d. = 8.1), and ranged from a low of 5 entries to a maximum of 82. Table 4 presents information on the nature of the first-tier activities from the ATUS for the total sample. The first part of the table shows the breakdown of activity occurrence for the 18 activities; the second portion shows information about the location of events. The final column of Table 4 displays the average duration of activities in minutes. A complete description of what these activity types entail can be found in Appendix B.

Table 4. Frequency of ATUS 2010 First Tier Behaviors

Behavior	Number of Behaviors	Percentage of Behaviors	Average Duration
Personal Activities	48006	18.74	206.11
Household Activities	36288	14.17	45.68
Care for HH Member	14408	5.63	30.35
Care for Non-HH Member	2952	1.15	39.95
Work	11791	4.60	177.79
Education	1616	0.63	133.41
Consumer Activity	8910	3.48	36.55
Professional/Personal Care Services	1311	0.51	45.45
HH Services	281	0.11	36.97
Gov. Services	96	0.04	51.47
Eating/Drinking	25818	10.08	34.69
Socializing/Leisure	40203	15.70	95.50
Sports/Exercise	3131	1.22	82.86
Religious	2211	0.86	79.35
Volunteer	1628	0.64	83.05
Telephone	2713	1.06	32.72
Traveling	51107	19.96	18.29
Error Codes	3635	1.42	70.18
Location			
At Home	107576	42.00	58.49
At Work	12021	4.69	152.70
In Transit	49809	19.45	19.20
Other Places	86699	33.85	140.86

The ATUS is intended to provide a picture of how people in general spend their time, and the observed pattern is expected to reflect the overall pattern in the population, whereas any single person-day would not necessarily be reflective of the frequency of events for that person. However, it is likely that there are differences across groups of people in patterns of time use. By grouping respondents with others having similar traits, the observed frequencies may more accurately reflect what is more typical for any given respondent, relative to the overall total. We grouped respondents based on four variables: age, divided into four categories (15-24, 25-44, 45-64, 65+); sex (2 categories); employed or not (2 categories); and whether there is child in the household or not (2 categories). Categorizing respondents based on all four variables jointly leads to 32 mutually exclusive and exhaustive divisions of the ATUS sample, and the frequency of first-tier behaviors are tallied for each of these divisions. The differences are at times stark. For example, care for a household member makes up 14.69% of activities for employed, 25-44 years old, women who have children in the household. By comparison, care for a household member makes up 5.63% of activities for the total sample, and 0.34% for not employed, 65+ years old, men who do not have children in the household. Many similar differences exist across the 32 groups. Given these differences, the 32 group frequencies, rather than the overall frequencies, will be used in the remaining analyses.

In order to estimate the effect of both response- and respondent- level characteristics on the likelihood that a recall failure occurs at any given period of time while filling out the diary, we used a multilevel logistic regression modeling approach , with the outcome being equal to 1 if a recall error occurs on a given entry, 0 otherwise. There are three levels to the model; the response level, nested within respondents, who in turn are nested within interviewers. Separate models are estimated for each additional set of characteristics at the different levels of hierarchy.

As a first step, a three-level random-intercepts only (i.e. null) model is estimated to calculate variance components and the intra-class correlation (ICC) coefficients. The next model includes only response characteristics, third is a model with response and respondent characteristics, and the final, full model includes response, respondent, and interviewer variables. Only cases used in the full model are used in all analyses. Although these models allow for better understanding of what each level adds to the model, unlike nested mixed-effect models using a continuous outcome, those using a categorical dependent variable are not strictly comparable (Bauer 2009; Fielding 2003, 2004; Hox 2010; Snijders and Bosker 1999).[3] Although many analyses compare these models, it is only wholly correct to do so when correcting the estimates using scalars (Enzmann and Kohler 2012; Hox 2010), which are not yet implemented in three-level models like those presented here.

At the response level, the amount of time elapsed (in minutes) between the event being recalled (as indicated by its time of day entry in the diary) and the taking of the survey (identified by the time of day the diary was taken) is included. Eighteen respondents did not have usable information about the start time of the survey and are not included in the analyses. The response characteristics of the entry immediately prior to the entry in consideration are also included. Although the ATUS time diary allows for recall and report in any order the respondent prefers, examining indicators of possible non-sequential reports, i.e. insertions or deletions of activities into the diary, show that 78.4% of respondents had zero insertions or deletions, suggesting the possibility that all these reports are sequential. Another 10.6% had 1, and another 5.0% had two, with 98.5% of these respondents having 5 or less insertions or deletions, suggesting that non-sequential recall occurs infrequently in the ATUS (Ruther et al. 2013). Therefore, the single lag

[3] This is due to the constant first level latent variance in categorical multilevel models; in logistic models the constant variance is set to $\pi^2/3$=3.29 (Hox 2010).

in the variables is selected as given the general forward sequential ordering of recall, the activity immediately prior will be the most recently activated in working memory.

Frequency of the immediately previous activity is indicated empirically, based on the actual frequency an activity occurred within each of the 32 divisions of the sample discussed previously, measured in percentages. Using the above example, employed, 25-44 years old women, with at least one child in the house would have a greater frequency value for care of a household member (=14.59) than not employed, 65+ years old, men who do not have children in the household (=0.34). Duration is measured as the reported number of minutes of previous activity. Besides the inclusion of duration, the interaction between frequency and duration is included as well. This interaction is included because although it is expected that more frequent activities generally will provide worse cues, longer frequently occurring events may be more distinctive due to the amount of time they took or the amount of episodic detail they include. An indicator is also included for whether one or more people were present during the previous activity, as are indicators for whether the previous activity occurred at home, at work, or while in transit. The impact of being at these places is estimated compared to the baseline category of all other places (e.g. others' homes, shopping centers, etc.).

The respondent-level characteristics used are those reported in Table 2. Whites are compared to all other races, and those with less than a high school degree and the unemployed are reference categories for those variables. In addition, the number of activities reported is included, as varying numbers of activities to report may affect the ability to recall all parts of the day. Also included is whether the respondent was reporting about a weekend as well as the estimated probability of response to the survey. Given that responses are clustered within respondents which in turn are clustered within interviewers, interviewers are included as the third

level of the model. Although recall failure comes from the respondent, interviewers have been shown to affect recall through their actions, such as probing, especially for calendar –type interviews (Belli et al. 2013). The ATUS does not include any interviewer characteristics, only a unique interviewer identifier, but measures of interviewer experience with the 2010 ATUS and overall error rates (other than recall errors) are included. The results of the four nested models are included in Table 5.

Table 5. Multilevel Estimations of Odds Ratio for Recall Failure

	(1) Null Model	(2) Response Characteristics	(3)Respondent Characteristics	(4) Interviewer Characteristics
Response Characteristics				
Elapsed Time		1.001*	1.001*	1.001*
Frequency		1.059*	1.048*	1.050*
Duration		1.002	1.002	1.002
Frequency*Duration		0.999*	0.999*	0.999*
Alone		1.649*	1.422*	1.438*
At Home		2.224*	2.182*	2.259*
At Work		0.054*	0.058*	0.059*
In Transit		0.303*	0.371*	0.367*
Respondent Characteristics				
Age			1.015*	1.017*
Weekly Income			1.000	1.000
Activities Reported			1.000	0.996
Female			0.994	0.998
White			0.970	0.937
Hispanic			0.801	0.959
Employed			0.808	0.822
Out of Labor Force			0.937	0.953
High School			0.862	0.791
Undergraduate			0.820	0.766
Graduate			0.641*	0.577*
Partner			1.000	1.034
Weekend			1.125	1.075
Response Propensity			0.577*	0.582*
Interviewer Characteristics				
Experience				1.000
Error Rate				0.238*
Random-effects Parameters				
Respondent Variance	0.870	0.821	0.740	0.730
Interviewer Variance	0.749	0.786	0.817	0.710
$\chi^2$of LR-Test against previous model		398.504*	72.724*	8.559*

Responses = 255,834 Respondents = 13,161 Interviewers = 69 *p<0.05

The random-effect variance estimates from the null model show that respondents differed in propensity to have a recall failure and interviewers impacted the likelihood of a recall failure differentially after controlling for respondent differences. The calculated ICCs from the null

model suggest that respondents account for 33% and interviewers 15% of the variability in the observed recall failures. Although the variance components are not strictly comparable, the log-likelihoods can still be used to assess model fit (Hox 2010). As shown in the final row of Table 5, each subsequent model improves model fit over the previous one. The full model, including interviewer characteristics, improves model fit over the model including response and respondent characteristics,$\chi_2^2 = 8.559$, p=0.014, and is the model selected for further analysis. The interviewer variance components vary across models, as may happen in categorical multilevel models (Enzmann and Kohler 2012), but it is worth noting that the smallest variances for both higher levels are for the full model.

The response-level effects show that, even with the short reference period, the effect of elapsed time is significant, with greater elapsed time between the activity and the survey increasing the chances of a recall failure. This effect of time is consistent with findings showing that the ability to recall autobiographical memory declines at farther points in the reference period (Thompson et al. 1996) and with the suggestion that in time diaries, memorable events that happen earlier in the day will not be remembered (Hektner et al. 2007). Using a different specification, including only the start time of the activities within the reported day instead of elapsed time from the interview found the same effect. Taken together, these findings do suggest the importance of time in recall, even over a relatively short period. It could also be that certain types of activities that are more likely to be forgotten also tend to occur earlier in the day. While more frequent activities may also be more likely to occur at certain times of the day, the inclusion of the frequency measure in the current model controls for this effect.

Importantly, the main effect of frequency of events has a clear effect in the expected direction. Events that occur more frequently are significantly more likely to be followed by a

recall failure, suggesting that previous activities in the sequence do provide cues to for the activity currently being recalled, and the frequency of the activity affects its quality as a cue. For each additional increase percentage point of frequency, the odds are 1.050 times greater that the next activity will be a recall failure. That more frequent behaviors are less effective cues are also consistent with findings that typical behaviors are more likely to be problematic in recall (Thompson et al. 1996), more likely to rely on semantic information (Linton 1982, Blair and Burton 1987), and be less accessible in memory (Brown 1997). The interaction between frequency and whether the day reported on was a weekend was not significant (not shown) and not included in the final model, suggesting that frequent events are less effective cues for differently structured days.

The main effect for duration is not significant; however, the interaction between duration and frequency is statistically significant. The coefficient indicates that as the duration of more frequently occurring events increases, their efficacy as cues also increases. Taking the main effect and the interaction as a whole suggests that frequently occurring events in a chronology are more likely followed by a recall error, but this likelihood is lessened the longer the event lasts. Conversely, for shorter durations, the probability of a recall failure increases at a more rapid rate as the frequency of an activity increases.

Figure 1 shows the impact of frequency at four different durations (1, 30, 60, and 120 minutes) and the interaction between these two on the predicted probability of a recall error (all other variables are held constant). Regardless of duration, as frequency increases, so does the probability of a recall error: lower probabilities of a following recall error are estimated for low frequency events. As indicated by the interaction, the probability of a recall error is relatively higher at higher durations for low frequency events and relatively lower for higher durations at

high frequency events. These findings are consistent with the reasoning outlined previously, that longer, frequent events become more distinctive in memory, attenuating the effect of frequency.

Figure 1. Predicted Probability of Recall Error Across Frequency for Four Durations

[FIGURE 1 HERE]

The remaining indicators at the response level add further evidence that characteristics of the immediate prior activity can affect recall of the target activity. When the previous activity occurs in transit or at work, the likelihood of a recall failure for the following event is significantly decreased relative to other locations. Conversely, being at home increases the likelihood of a following recall failure compared to events at other places. Similarly, being alone while doing something is more likely to be followed by a recall failure than when an event is conducted with others present. Both being at home and being alone are conceivably less distinctive or salient cues than being other places or being with others. While being at work may be a common event, being less distinctive generally, it may be useful for placing events in the chronological sequence. Respondents may use a metastrategy to recall events, relying on episodic information in some instances and generic information in others, such as when the added cognitive effort is unlikely to yield locating an activity in memory (Brown 2008). It may be that being at work is an effective memory because it has generally scheduled beginning and end points. Again, the impact of being in transit may in part be attributed to the fact that if the respondent fails to mention travelling, the interviewer is instructed to probe to see if there was a mistake. In these cases, the added probing by the interviewer may also affect the recall of the following event.

Most of the respondent-level variables have non-significant effects on the predicted likelihood of a recall failure, although a few do in ways consistent with expectations. Those with higher response propensity are less likely to have recall errors, consistent with the notion that those who are more resistant initially expend less cognitive effort if they do respond (Fricker 2007; Olson 2006). Age and education also have significant impacts on the likelihood of a recall failure, with older and less educated (with undergraduate degree approaching significance, p=0.057) respondents having a higher likelihood of a recall failure; in general, those with lower cognitive ability are more like to make recall and nonresponse errors (Schwarz et al. 1999; de Leeuw et al. 2003).

Including interviewer measures improves model fit, but only one measure is significant. The interviewer's overall error rate (other than recall failures) is significantly related to recall failure, but not their experience. The effect of overall error rate is opposite to expectation, with higher error rates by the interviewer leading to lower probabilities of recall failure. As noted, the remaining errors are more likely to be related to both the interviewer and respondent. The different directional effect of other errors suggests that the characteristic driving certain types of errors are uniquely different, and in some way opposite, of those in aiding recall of the respondent. A similar finding was reported by Belli et al. (2013), who show that specific types of interviewer and respondent interactive behaviors will assist or detract from recall accuracy, depending on the characteristics of the event being recalled.

4. Discussion

Understanding recall in general and in the survey context is important given the frequency of questions asking about autobiographical information. The importance of recall is

heightened to an extent in diary and time use surveys as the main purpose of these surveys is the compilation of autobiographical events. The current study focuses on recall and recall failures specifically in one of the largest time use diary surveys, the American Time Use Survey. Further, the results presented here may be informative to survey methods in general, as we believe it sheds light on memory processes in responding to autobiographical memory questions generally.

The basis for the model presented here is that recall failures are influenced by the entire survey process, not just the characteristics or abilities of the respondent and the interviewer. The current research adds to a growing literature showing the impact that item- or response-level characteristics have on outcomes in conjunction with respondent and interviewer effects, suggesting the need for a multilevel approach in survey analysis. The importance of response-level characteristics is highlighted by the significant effects found at this level, whereas far fewer respondent characteristics are significant. Once other aspects of the survey process are accounted for, the impact of respondent characteristics may be less than previously believed. That is not to say that respondents do not matter - they very much do. The variance components of the models show there is still a substantial portion of variance remaining relating to respondents, even after controlling for a number of response and respondent characteristics. Further, there are some significant respondent effects for variables consistently used as proxies for cognitive ability or effort.

Given the short reference period, recall failure seems less likely in the ATUS than in other surveys. However, the events furthest away in time are still less likely to be recalled. The impact of cognitive ability at the respondent-level and retention interval at the response level underscore the importance of memory. The characteristics of prior activity included in the model seem to affect memory as well, an impact that we believe reflects cuing. Our results

indicate that more frequently occurring events are less effective cues in assisting sequential recall. This result is in line with findings suggesting that recall of more frequently occurring behaviors rely on semantic rather than episodic information (Linton 1982, Blair and Burton 1987), and that frequently occurring events are not as accessible as unique ones (Brown 1997) and provide less effective cues.

The effect of frequency is modulated by the duration of the previous event. Although duration does not appear to have a significant direct impact on recall failures, the longer a frequent activity, the better it is as a cue. The lack of a main effect for duration in combination with the significant interaction effect indicates that high frequency activities become more effective cues only when the activity is long, likely making such activities considerably more distinctive from high frequency events of shorter duration. Other characteristics including the who and where of an event also influenced recall of the next activity; these been found elsewhere important as cues in autobiographical memory, albeit for the target event as opposed to the following event (Wagenaar 1986, Thompson et al. 1996). These results are robust. We ran a similar model using 2008 ATUS data and found nearly identical results in regards to significance and direction of effects.

What are the implications of our results for future design? If a respondent recalls an event that is less useful in aiding recalling the next event, it is impossible to change the event itself, but additional aspects of the event may be leveraged with alternative cuing methods. When a respondent fails to recall an event, additional questions about the details of the forgotten event may be fruitless – it is forgotten, after all. However, additional questions can then be asked about that prior activity in an attempt to add to the detail and value that memory as a cue for the target event. As an example, for more frequently occurring events, questions about how the previous

event was unique compared to other events of the same type may increase its usefulness as a cue. While this may violate standardization, some diaries such as the ATUS do not use standardized survey interviewing.

There are limitations to our research. First, the data come from a very specific survey design, albeit one of the largest time diaries collected. It is a time use diary collected by interviewers over the telephone encompassing a reference period of only the previous day, and different designs and recall periods may lead to differing results. Further, the ATUS diary promotes forward chronological recall, which may result in different recall effects than if greater variability was observed in the order of retrieval. However, even in this case, the immediate temporal link should still exist (if asking about a sequence of events), and information should remain in working memory. We believe that these findings are consistent with the bulk of literature on memory and recall, and suggest that these represent outcomes from cognitive processes that would occur in any number of autobiographical questions in surveys.

Second, we are only able to examine observed recall failures as coded within the ATUS. Although this measure does indicate a failure occurred, it only measures respondents' admission to forgetting, as opposed to erroneous reports of other activities. Finally, there is some limitation to the variables used in the model. Although it would be ideal to know the frequency of events in each respondent's life, the design of the ATUS does not allow for this, capturing only one day per person. However, we believe that by categorizing respondents into 32 subgroups adequately represents the frequency of occurrence for people with given characteristics. Still, considerable respondent variation doubtless still remains. Finally, no characteristics are available for interviewer, other than observed survey outcomes, and these characteristics may be important in understanding why errors are made.

Future research could focus on some of these limitations, including studies on how diary design affects recall strategies and how this influences the use of prior responses as recall cues. Audit trails (paradata) would also be useful in better understanding the response process more fully and could allow additional data quality measures (e.g. Ruther et al. 2013). Research should also examine other survey designs and the differences between self-administered and interviewer-administered surveys, while collecting more interviewer characteristics. These additional variables may explain the finding that higher interviewer rates of other types of error are related to lower probabilities of a recall failure.

References:

Abraham, K. G., Helms, S., & Presser, S. (2009). How Social Processes Distort Measurement: The Impact of Survey Nonresponse on Estimates of Volunteer Work in the United States. *American Journal of Sociology, 114*, 1129-1165.

Bauer, D.J. (2009). A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika, 74*, 97-105

Barsalou, L. W. (1988). The Content and Organization of Autobiographical Memories. In U. Neisser & E. Winograd (Eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, 193-243. Cambridge: Cambridge University Press.

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, 383-406.

Belli, R. F., Agrawal, S., & Bilgen, I. (2012). Health Status and Disability Comparisons between CATI Calendar and Conventional Questionnaire Instruments. *Quality and Quantity*, 46, 813-828.

Belli, R. F., Alwin, D. F., & Stafford, F. P. (2009). The Application of Calendar and Time Diary Methods in the Collection of Life Course Data. In R. F. Belli, F. P. Stafford, & D. F. Alwin (Eds.), *Calendar and Time Diary methods in Life Course Research* (pp. 1-12). Thousands Oaks, CA: Sage.

Belli, R. F., Bilgen, I., & Al Baghal, T. (2013). Memory, Communication, and Data Quality in Calendar Interviews. *Public Opinion Quarterly, 77,* 194-219.

Belli, R. F., & Callegaro, M. (2009). The Emergence of Calendar Interviewing: A Theoretical and Empirical Rationale. In R. F. Belli, F. P. Stafford, & D. F. Alwin (Eds.), *Calendar and Time Diary methods in Life Course Research* (pp. 31-52). Thousands Oaks, CA: Sage.

Blair, E., & Burton, S. (1987). Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Question *Journal of Consumer Research, 14,* 280-288

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing Life as it is Lived. *Annual Review of Psychology, 54,* 579-616.

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

Brewer, W. F. (1994). Autobiographical Memory and Survey Research . In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective reports* (pp. 11-20). New York: Springer-Verlag

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1539-1553.

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 898-914.

Brown, N. R (2008). How metastrategic considerations influence the selection of frequency estimation strategies. *Journal of Memory and Language, 58,* 3-18.

Burton, S., & Blair, E. (1991). Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys *Public Opinion Quarterly, 55,* 50-79

Conrad, F., Brown, N. R., & Cashman, E. (1998). Strategies for estimating behavioral frequency in survey interviews. *Memory, 6,* 339-366.

Couper, M. P., & Kreuter, F. (2013). Using Paradata to Explore Item Level Response Times in Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 271-286.

Conway, M. A. (1996). Autobiographical Memories and Autobiographical Knowledge. In D. C. Rubin (Ed.), *Remembering Our Past: Studies in Autobiographical Memory* (pp. 67–93). Cambridge: Cambridge University Press.

De Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics, 19,* 153-176.

Enzmann, D. and Kohler, U. (2012) *Rescaling results of nonlinear probability models to compare regression coefficients or variance components across hierarchically nested models*. Paper presented at German Stata Users Group Meeting, Berlin.

Fielding, A. (2003). Ordered category responses and random effects in multilevel and other complex structures. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fielding, A. (2004). Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity, 38,* 425–433.

Freedman, V. A., Stafford, F., Conrad, F., Schwarz, N., & Cornman, J. C. (2012). Assessing Time Diary Quality for Older Couples: An Analysis of the Panel Study of Income Dynamics' Disability and Use of Time (DUST) Supplement. *Annals of Economics and Statistics, 105,* 271-289.

Fricker, S. (2007). *The Relationship between Response Propensity and Data Quality in the Current Population Survey and the American Time Use Survey.* (Unpublished Doctoral Dissertation). University of Maryland, Maryland.

Gershuny, J. (2000). *Changing Times: Work and Leisure in Postindustrial Society*. New York: Oxford University Press, Inc.

Harvey, A. and Royal, M. (2000) *Use of Context in Time-use Research*, paper presented at Expert Group Meeting on Methods for Conducting Time-Use Surveys, New York, 23–27 October

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life*. Thousand Oaks, CA: Sage Publications, Inc.

Hox, J.J. (2010). *Multilevel Analysis.* New York, NY. Routledge

Juster, F. T. (1985). The Validity and Quality of Time Use Estimates Obtained from recall Diaries. In F.T. Juster and F.P. Stafford, (Eds.), *Time, Goods, and Well-being* (pp. 63–91). Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Juster, F. T., & Stafford, F. P. (1991). The Allocation of Time: Empirical Findings, Behavioral Models, and Problems of Measurement. *Journal of Economic Literature*, *29*, 471-522.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, *306*, 1776-1780.

Linton, M. (1982) Transformations of memory in everyday life. In Neisser, U. (ed.) *Memory Observed: Remembering in Natural Contexts*, San Francisco, Freeman.

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

Menon, G. (1993). The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies, *Journal of Consumer Research*, 20, 431-440.

Michelson, W. M. (2005). *Time Use: Expanding Explanation in the Social Sciences*. Boulder, CO: Paradigm Publishers.

Neter, J., & Waksberg, J. (1964). A Study of Response Errors in Expenditures Data from Household Interviews. *Journal of the American Statistical Association, 59,* 18-55.

Olson, K. M. 2006. Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly* 70:737-58.

Olson, K.M. & Bilgen, I. 2011. The Role of Interviewer Experience on Acquiescence *Public Opinion Quarterly* 75:99-114

Olson, K. M. & Peytchev, A. 2007. Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes *Public Opinion Quarterly* 71:273-286.

Pentland, W. E. (Ed.). (1999). *Time Use Research in the Social Sciences*. New York: Plenum Publishing Corporation.

Phillips, A.L., Al Baghal, T., and R.F. Belli. (2013, May). *Troubles with Time-Use: Examining Potential Indicators of Error in the American Time Use Survey.* Paper presented at AAPOR 68th Annual Conference, Boston, MA.

Pillemer, D. B., Goldsmith, L. R., Panter, A. T., & White, S. H. (1988). Very Long-Term Memories of the First Year in College. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14, 709-715.

Robinson, J. A. (1986). Temporal Reference Systems and Autobiographical Memory. In D. C. Rubin (Ed.), *Autobiographical Memory* (pp. 159-188). New York: Cambridge University Press.

Robinson, J.P. (1985) The Validity and Reliability of Diaries versus Alternative Time Use Measures. In F.T. Juster and F.P. Stafford, (Eds.), *Time, Goods, and Well-being* (pp. 33–62). Ann Arbor, MI: Institute for Social Research, The University of Michigan.

Schwarz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation.* Mahwah, New Jersey.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of Retrieval as Information: Another Look at the Availability Heuristic. *Journal of Personality and Social psychology*, *61*, 195.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, *49*, 388-395.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (Eds.). (1999). *Aging, Cognition, and Self-reports.* Washington, D.C.: Psychology Press.

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Sage.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stafford, F. (2009). Timeline Data Collection and Analysis: Time Diary and Event History Calendar Methods. In R. F. Belli, F. P. Stafford, & D. F. Alwin (Eds.), *Calendar and Time Diary methods in Life Course Research* (pp.13-30). Thousands Oaks, CA: Sage.

Sturgis, P. (2004). The Effect of Coding Error on Time Use Surveys Estimates. *Journal of Official Statistics*, *20*, 467-480.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass.

Thompson, C.P., Skowronski, J.J., Larsen, S.F., & Betz, A.L. (1996). *Autobiographical Memory: Remembering What and Remembering When.* Hillsdale, NJ: Erlbaum.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Ver Ploeg, M., Altonji, J., Bradburn, N., DaVanzo, J., Nordhaus, W., & Samaniego, F. (Eds.). (2000). *Time-use Measurement and Research: Report of a Workshop*. Washington, DC: National Academy Press.

Wagenaar, W. A. (1986). My Memory: A Study of Autobiographical Memory Over Six Years. *Cognitive Psychology*, *18*, 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*, 51-68.