Contrasting Stylized Questions of Sleep with Diary Measures from the American Time Use Survey

International Conference on Questionnaire Design, Development, Evaluation and Testing (QDET2)

November, 2016

Robin L. Kaplan, Brandon Kopp, and Polly Phipps

Office of Survey Methods Research

Bureau of Labor Statistics[1]

[1] Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

Abstract

In the American Time Use Survey (ATUS), interviewers use a set of open-ended questions to walk respondents chronologically through their activities during the prior 24-hours. In contrast, other surveys ask people about "the average, normal, or typical" time spent on activities (stylized questions). Estimates of sleep duration in the ATUS and other diary measures exceed those of stylized questions by approximately 1.7 hours – termed the sleep gap. Our research draws on a variety of evaluation methods (behavior coding, cognitive interviews, quantitative research, and a validation study using sensor data) to examine reasons for the discrepancy between diary and stylized sleep measures and to uncover potential sources of measurement error that may contribute to the sleep gap. We discuss the strengths and weaknesses of each method and how they can build off one another in the questionnaire evaluation process to gain a deeper understanding of a substantive survey methods issue.

Keywords: Time use, stylized questions, behavior coding, cognitive interviews, sensor data, sleep measures, measurement error

1. Introduction

Researchers, government agencies, and health institutes have become increasingly interested in collecting data on how people spend their time, as time use can have important economic, health, and policy implications. Two common methods of collecting time use data are time diaries and stylized questions, both of which can be interviewer or self-administered. Time diaries involve prompted recall, where respondents report on all of their activities for a specified period of time, such as the previous 24 hours. Stylized questions ask respondents to report the amount of time they spend on different activities on an *average*, *typical*, or *usual* day or week (e.g., "How many hours do you work on a typical day?").

Although both methods collect data on time use, diary measures are typically considered more valid and reliable than stylized measures (Juster, Ono, & Stafford, 2003; Kan & Pudney, 2008), as they focus on a set reference period and are less prone to recall and estimation bias. However, diary methods are more expensive to administer and burdensome for respondents to complete compared to stylized measures (Schulz & Grunow, 2012). Diary and stylized measures also produce different time use estimates across a variety of activities (Kan & Pudney, 2008). For instance, researchers have observed a "gap" in estimates of self-reported paid work hours using stylized questions and diary measures, with estimates from stylized questions exceeding diary estimates by an average of 3.52 hours per week (Lin, 2012). A similar pattern is found for hours spent on household chores, where stylized estimates exceed diary estimates by about 0.79 hours for women and 1.96 hours for men (Kan, 2008). Similar gaps occur with a wide range of other activities, including religious service attendance (Brenner, 2011), exercise (Adams et al., 2005), and sleep (Miller et al., 2015) – the topic of this chapter.

1.1 The Sleep Gap

Unlike other activities, for sleep, diary measures tend to exceed stylized measures. For instance, the American Time Use Survey (ATUS), which measures all activities (including sleep)

a respondent did on the previous day, found in 2014 that Americans aged 18 years and over slept 8.7 hours per night on average. In contrast, other national U.S. surveys (e.g., the National Health Interview Survey; NHIS) use stylized questions to collect data on how many hours per night people sleep such as, "On average, how many hours of sleep do you get in a 24-hour period?" Surveys using stylized questions consistently find that people report sleeping between 6.9 to 7.1 hours per night (Ford, Cunningham, & Croft, 2015). This constitutes roughly a 1.7 hour gap between diary (ATUS) and stylized measures of sleep duration. Despite these surveys being nationally representative with similar sampling methodologies, they produce different estimates of sleep duration. These differences may therefore be partly due to the way questions about sleep are asked.

1.1.1 Diary Measures of Sleep in the American Time Use Survey

The ATUS measures how Americans allocate their time in a one-day time frame using a sample of approximately 26,000 people each year (Phipps & Vernon, 2009). One individual from each sample household (aged 15 years or older) is selected to respond, and he or she participates in a computer-assisted telephone interview (CATI). The interviewer asks the respondent about what he or she did over a 24 hour period from 4:00 a.m. on the diary day until 4:00 a.m. on the interview day. Each activity is recorded along with either the duration or the start and stop times for the activity. For survey estimates, the total duration of time that people spent doing various activities is calculated.

In addition to the ATUS, other time diary surveys have shown that Americans 18 years and older report sleeping an average of 7.7 hours per night (Hale, 2005), and 8.1 hours per night (Biddle & Hamermesh, 1990). Statistics Canada's General Social Survey, which uses a similar methodology to the ATUS, found that adults over 15 years old reported mean sleep durations between 8.0 and 8.3 hours (Hurst, 2008). Across multiple studies time-use methodologies tend to yield sleep estimates that are eight hours or longer in duration.

1.1.2 Stylized Measures of Sleep

Unlike diary measures, stylized questions ask respondents directly about the amount of sleep they get in a typical, usual, or average day. Stylized questions can also differ in terms of the time frame they

ask about, ranging from the previous day, a typical day, or a typical week. For example, the National Center for Health Statistics (NCHS) collects sleep duration data via stylized questions. Two of their surveys, the NHIS and the Behavioral Risk Factor Surveillance System (BRFSS) are both conducted with samples of adults, aged 18 years or over. Both surveys ask respondents, "On average, how many hours of sleep do you get in a 24-hour period?" Responses are recorded as integer values (i.e., decimal or fractional reports are rounded to the nearest whole hour). Respondents in the 2014 NHIS and BRFSS reported an average of 7.1 hours of sleep per 24-hour period; the median amount of sleep respondents reported was 7.0 hours for both surveys.[2] Thus, national surveys using stylized sleep questions produce lower sleep duration estimates than those of the ATUS and other diary measures.

1.1.3 Response Processes in Self-Reported Sleep

When respondents answer survey questions, they often follow a four-step process where they try to understand (comprehend) the survey question, retrieve the relevant information to arrive at an answer, make a judgment (e.g., calculate an average), and finally give a response (Tourangeau, 1984). Measurement error can occur at any stage of the survey response process. Below, we hypothesize below how diary and stylized measures may affect reports of sleep at each stage of the response process and contribute to the sleep gap.

Comprehension. How respondents define "sleep" may affect how they report on their sleep. Some respondents may include naps, resting with their eyes closed, dozing off, or trying to fall asleep, while others may not. Some may interpret stylized questions as asking only about continuous episodes of nighttime sleep, excluding naps or times they were awake at night (Canfield et al., 2003). Since diaries ask respondents to report on a full day of activities, they are more likely to capture daytime naps. Thus, diary measures may capture additional sleep that a

[2] There is no equivalent question for weekend nights, which typically show longer sleep durations, so this number is likely a low estimate of respondents' total sleep.

stylized question might not. Survey context may also play a role – a survey about health may cause respondents to interpret the term "sleep" differently than a general survey about time use.

Recall. Both time diaries and stylized questions may be prone to recall error. Time diaries rely on rely on respondents' ability to recall the activities they did the previous day. It may be difficult for some for some respondents to recall the precise times they fell asleep and woke up. In contrast, stylized questions require respondents to search their memories for a representative set of days that reflect their typical sleep pattern, adjusting for weekends, holidays, or other events that may have affected their usual sleep (Kan & Pudney, 2008).

Judgment. Both diary and stylized sleep measures may be prone to measurement error if respondents cannot directly recall information that would help them report on sleep, and estimate it instead. For time diaries, respondents may rely on their typical routine to infer what time they must have fallen asleep and woken up. In contrast, stylized questions require respondents to make a judgment about the typical or average amount they slept during that period (Kan & Pudney, 2008). This judgment requires respondents to use an estimation strategy (e.g., rate retrieval, rate and adjustment, averaging; Conrad, Brown, & Cashman, 1998). Estimation strategies are prone to systematic biases, such as rounding or calculation errors.

Response. The last stage is reporting a response, which can be prone to filtering and social desirability concerns. Respondents may also edit their answers differently depending on the survey topic (e.g., Couper, Conrad, & Tourangeau, 2007). Diary measures of sleep are usually collected within the context of time use activities, whereas stylized questions ask directly about sleep, which may affect respondents' answers (Schwarz, Strack, & Mai, 1991), for example, by encouraging them to report on norms and beliefs about the appropriate amount one should sleep rather than their actual behaviors (Bonke, 2005). This difference in question context may contribute to observed differences in sleep estimates. Furthermore, it has been suggested that stylized questions may be more prone to errors arising from respondents' editing or rounding their answers than diary measures (Kan & Pudney, 2008).

Aside from the survey response process, another potential contributor to the sleep gap is the way in which a survey defines and measures sleep duration. ATUS includes napping, falling asleep, and sleeplessness in its sleep estimate. Other surveys using stylized questions tend to define sleep as the longest continuous episode of sleep (Silva et al., 2007). It takes the average American about 20 minutes to fall asleep (Silva et al., 2007), but if respondents report falling asleep within 30 minutes of going to bed, the ATUS records that time as sleep, which may inflate ATUS sleep estimates. In addition, NCHS records sleep duration as integer values (i.e., decimal or fractional reports are rounded to the nearest whole hour), which may affect the distribution of responses.

1.2 The Present Research

This research drew on a range of questionnaire evaluation methods (i.e., behavior coding, qualitative interviews, a quantitative survey, and a validation study) to assess possible reasons for the discrepancy between diary and stylized sleep measures. We explored the cognitive processes involved in answering diary and stylized sleep questions to identify possible sources of measurement error in both measures. We used a sequential mixed research approach (Creswell & Creswell, 2017) to garner unique insights from each method, to build off the previous research findings, and to contribute in different ways to our understanding of the sleep gap and measurement error.

Study 1 involved behavior coding of ATUS interview transcripts to investigate issues related to interviewer and respondent behaviors that may affect the ATUS sleep estimates. Behavior coding allows researchers to identify concepts and tasks that respondents and interviewers may struggle with and detect practices that could be associated with survey measurement error (e.g., Van der Zouwen & Smit, 2004; Dykema, Lepkowski, & Blixt, 1997). Drawing on the findings from the behavior coding research, in Study 2 we carried out cognitive interviews to uncover more about respondents' cognitive processes when answering diary and stylized questions about sleep. Cognitive interviews provide an in-depth understanding of a

respondent's thought processes and reactions to a question, and can reveal the content validity of a question (i.e., does the question measure what it is intended to measure) and what possible sources of error may underlie the question (Willis, 2005). Findings from the cognitive interviews were used to generate a set of hypotheses for Study 3, which involved a quantitative approach. We tested hypotheses about diary and stylized measures, survey context, and providing definitions of sleep, using a large online sample to make statistical comparisons across experimental groups. Finally, in Study 4, we carried out a validation study comparing self-reported sleep (diary and stylized measures) against sensor data obtained from devices worn by study participants for one week that tracked their activity level, including their sleep.

In the following sections, we will describe each study and its findings, the methods used, and how each method revealed different potential sources of measurement error associated with diary and stylized sleep measures that may help explain the sleep gap.

1.2 Study 1: Behavior Coding

Study 1 used behavior coding, a questionnaire evaluation method in which researchers systematically code interviewer and respondent interactions during the survey interview, either 'live' during the interview or from an audio recording of the interaction (e.g., Van der Zouwen & Smit, 2004; Dykema et al., 1997). We started with behavior coding of ATUS interview transcripts to gain a deeper understanding of issues related to interviewer and respondent behaviors that may affect the ATUS sleep estimates.

Sample and Demographics. A total of 104 ATUS interviews conducted by 36 interviewers during 2008 were audio-recorded with respondent consent. Interviewers recorded up to three consecutive interviews each. The demographics of the study sample were compared to the full 2008 ATUS sample and no significant differences were found for sex, age, race, and other demographic variables (p-values > 0.05). Thus, while the interviews were not randomly selected, the study sample can be viewed as demographically representative of the full ATUS sample (Denton, Edgar, Fricker, & Phipps, 2012).

Coding Scheme Development. A team of survey methodologists developed the coding scheme, coded the interview transcripts, and analyzed the data. The main unit of analysis was a sleep episode, defined as a full conversation around sleep (e.g., the first and last conversational turn related to sleep). Within a sleep episode, several items were coded: the type of sleep (sleep or nap); mentions of sleep time, which consisted of reported wake times, sleep times, or sleep duration; and whether the respondent used a recall strategy (e.g., alarm clock) or qualifier when providing time (e.g., about, around). We also coded interviewer behaviors (e.g., use of scripted versus unscripted probes).

For reliability purposes, 12 transcripts were double coded, or 11.5% of the total transcripts. For quantitative variables (e.g., sleep duration), percent agreement[3] was calculated, and ranged from 0.85 to 1.00. For categorical variables, kappas were calculated and resulted in moderate to almost perfect agreement, ranging from 0.45[4] to 1.0.

1.3.1 Behavior Coding Results

Sleep Episodes. Respondents reported an average of 2.2 episodes of sleep (*SD* = 0.57, range = 1 to 4) during the 24-hour period covered by the diary. These sleep episodes were most often found to occur at the beginning and end of the diary day. Respondents often reported the time they woke up and fell asleep on the diary day and the time they woke up on the day of the interview.

Sleep Duration. On average, respondents reported sleeping 8.53 hours (*SD* = 1.98; range = 2.42 to 15.48 hours) per 24-hour period. Just over a fifth (22.0%) of respondents reported at least one nap. Of the respondents who reported taking at least one nap, the total time spent napping was 1.35 hours on average (*SD* = 0.80 hours). These sleep estimates were comparable to the ATUS published sleep estimates.

[3] Percent agreement was used because it is a measure of inter-rater reliability between two coders using quantitative variables (see McHugh, 2012). Pearson *r* correlations were similar and ranged from 0.89 to 1.00.

[4] Only one category, respondent qualifications about their sleep, had a kappa of 0.45, or moderate agreement. The remaining kappas were all 0.85 or above. Qualifications may have been more difficult to code due to ambiguity in the language respondents used to describe their activities and that it was sometimes unclear if the qualification, or multiple qualifications within a sleep episode, referred to sleep or other pre-sleep activities.

Interviewer/Respondent Interactions. We analyzed the number of conversational turns from when a respondent first mentioned sleep until the final sleep-related turn. The number of turns gives an indication of how complicated the process of recording sleep can be. In an average interview, 21 interviewer-respondent interactions were needed to record sleep, amounting to 11.9% of the interactions in the interview. An average of 6.7 turns was needed to record the time of waking or falling asleep. Interestingly, the mean number of turns required to record the time that respondents fell asleep was much higher (M = 11; range = 1 to 52 turns) than the number of turns to record wake times (M = 3.9, range = 1 to 13), suggesting greater cognitive task complexity in reporting falling asleep times versus waking up times.

Wake Time Probes. At the beginning of the diary day, if a respondent reports that he or she was sleeping at 4 a.m., interviewers are trained to ask a non-leading question, such as "What time did you wake up?" Often, however, interviewers will rephrase this question in a potentially leading way. For example, an interviewer might ask "What time did you get up?" which could be interpreted as asking about the time they physically left their bed rather than when they woke up. Interviewers used leading question wording 68.9% of the time. The most common leading question was "What time did you get up?" with "What time did you wake up yesterday morning?" or "What time did you wake up this morning?" also being commonly asked. The latter is problematic, because it suggests to respondents that the interviewer is not interested in capturing times awake during the night.

"Went to Bed" Probes. Another place where the use of language may increase measurement error is the transition between wakefulness and sleep. Near bed time, respondents may use phrases such as "I went to bed." This could mean when they went to sleep, but it could also mean when they laid in bed awake, e.g. watching TV, reading, or trying to fall asleep. This time should not be recorded as sleep. When respondents say they "went to bed" interviewers are trained to use the following scripted probes:

1. What time did you fall asleep?
2. Did you groom, read, watch television or something else before you fell asleep?

Respondents used the phrase "went to bed" in a total of 76 (or 73.0%) of interviews with interviewers probing respondents about this in 72 (or 95.0%) of instances. Of the probes used by interviewers, 42.0% were scripted and 53.0% unscripted. The most common unscripted probe was some variation of "Did you go to sleep immediately?"

Respondent Qualification of Answers. We coded whether respondents used qualifiers (e.g., about, around, or maybe) when reporting the time they woke up or fell asleep. The use of qualifiers gives an indication of whether the respondent was confident in their answer. Respondents averaged 1.8 qualifiers per interview across the average of 3.1 reports of falling asleep or waking. Overall, of the 56.9% of the time respondents reported a transition between sleep and wakefulness, they did so with some amount of reservation. Respondents were slightly more likely to qualify their responses when reporting the time they fell asleep (60.6% of the time) than times they woke up (49.3% of the time), indicating they may have had more difficulty recalling or estimating their sleep versus wake times.

Recall Strategies. To better understand how respondents formulate their response of when they woke up or went to sleep, any explicitly mentioned response strategy was coded. That is, if a respondent said "I always get up at 8 a.m.," their response strategy would be coded as "typical routine." If the respondent said "My alarm went off at 8 a.m.," their response strategy would be coded as "alarm clock." In 83 cases (25.1% of the total 331 sleep mentions across all interviews), respondents mentioned some type of response strategy. Table 1 shows the frequency of those strategies.

Table 1. Frequency of respondents' strategies when reporting sleep and wake times.

Response Strategy	**Uses**	**%**
Alarm	33	40.0
TV	32	38.6
Viewed clock	7	8.4
Direct Recall	0	0.0
Guess	0	0.0
Typical Routine	9	10.8
Other	2	2.4

The most common response strategies were reports of an alarm clock going off (usually to recall a wake time) or specific mentions of a TV show e.g., "The news was on so it must have been 11 p.m.," (usually to recall a sleep time).

Summary. Behavior coding provided insight into how sleep is reported and recorded in ATUS interviews. We found that the interactions, especially those surrounding when a person falls asleep, are often complex, indicating that respondents may have difficulty recalling or estimating their sleep time. Interviewers often used unscripted or leading probes when requesting this additional information. This could lead to underreporting of both time spent awake during the night and time spent lying awake in bed before physically getting up, which might inflate ATUS sleep estimates, and contribute to the sleep gap.

While the behavior coding study was useful to understand interviewer-respondent interactions in the ATUS, we did not have insight into how respondents answer questions about sleep. To expand on the knowledge garnered by the coded interactions, we brought participants to the lab to conduct cognitive interviews.

1.3 Study 2: Cognitive Interviews

Cognitive interviews provide an in-depth understanding of a respondent's thought processes and reactions to a question. Cognitive interviews can uncover the content validity of a question (i.e., does the question measure what it is intended to measure) and what possible sources of error may underlie the question (Willis, 2005). To follow-up on the results of the behavior coding, we conducted cognitive interviews to gain insight into how respondents report on their sleep. We explored differences between diary and stylized measures at each stage of the response process (comprehension, retrieval, judgment, and reporting) with the aim of identifying possible sources of measurement error that may contribute to the sleep gap.

The cognitive interviews for this study were conducted by BLS researchers using the ATUS interview protocol and lasted approximately one hour each. Participants were asked to complete an

abbreviated ATUS daily recall interview about the prior 24-hour period (from 4 a.m. to 4 a.m.) and answer a set of stylized questions about their sleep, as follows:

1. Diary measure: The total number of hours participants reported sleeping in the previous 24-hour period (the abbreviated ATUS interview measure).
2. General stylized measure: "How many hours do you sleep at night on an average weekday?"
3. Last week stylized measure: "Thinking about the past week, on average, how many hours did you sleep each night?"

Interviews took place Tuesday to Friday, meaning all the cognitive ATUS interviews covered a weekday.[5] The ATUS interview focused on times when participants were likely to have woken up and gone to sleep (i.e., 4 a.m. to 10 a.m. and 7 p.m. to 3:59 a.m.). Participants were also asked about any naps taken between 10 a.m. and 7 p.m.[6] The order of the ATUS interview and stylized questions was randomized. Following the administration of the survey questions, participants answered retrospective probes aimed at understanding how they arrived at their answers to each of the three questions.

Participants. We recruited 29 participants (11 male, 18 female) from the Washington, DC metro area. The mean age was 46.0 (*SD* = 14.1), with a range of 21 to 69 years old. Nine participants had a high school diploma or equivalent, six had some college, seven had a college degree (Associate's/Bachelor's), and five had an advanced degree (Master's/Doctorate).

1.4.1 Cognitive Interview Results

Comprehension. First participants were asked to describe what the word "sleep" meant to them, what activities they included as part of sleep, and whether these activities were included in their answers. Responses varied from narrow definitions (e.g., being fully unconscious) to broader ones (e.g., dozing off, trying to fall asleep). As seen in Table 2, participants were fairly evenly divided between those using a narrow or broad definition of sleep, with those having a

[5] People tend to get more sleep on weekends, so we limited the study to weekdays only (Ford & Cunningham, 2015).
[6] Results did not change significantly whether including or excluding naps in the ATUS sleep measures.

broad definition reporting sleeping approximately one hour more on average across each of the three measures than those with a narrow definition of sleep. .

Table 2. Mean sleep duration across varying participant definitions of sleep

	Diary	**General stylized**	**Last week stylized**
Narrow sleep definition (*n* = 15)	7.25	6.90	5.92
Broad sleep definition (*n* = 13)	8.18	7.15	6.93

Diary Recall. While most participants could confidently recall what time they woke up in the ATUS interview because they followed a structured schedule and set an alarm for the same time each morning, the majority could not directly recall what time they fell asleep. Many used the TV program they were watching that evening to infer what time they must have fallen asleep. Some looked at the clock or guessed, and only two knew because they have a regular, scheduled bedtime.

Stylized Recall/Estimation. Participants were asked to describe how they answered the last week stylized sleep question. Responses fell into the following categories:

- Recalled directly: Participants reported having a structured schedule and knew what time they fell asleep and woke up (e.g., "I went to bed at about the same time every night. I knew I went to bed at the same time; I have the same schedule.")
- Rate retrieval: Participants recalled the typical number of hours they sleep in a night, and used that as the average (e.g., "My usual hours of sleep are between 11 and 6 for the work week.")
- Rate and adjustment: Participants recalled the typical number of hours they slept in the past week, and then made adjustments for events that happened that week (e.g., "Since it was a long weekend, that came to my mind. Thought of the average and then subtracted a bit because yesterday was busier than usual.")
- Calculation: Participants summed the number of hours slept each night that week, and then took the average (e.g., "Tried to apply a median. Some nights that were shorter, some were greater. 4 + 3 + … and came up with actual average.")

- Estimate/Guess: Participants could not recall how much they slept in the past week, so they estimated or guessed (e.g., "I took a guess. I went by the activities I was doing last week.")

Table 3 shows the frequency of participants using the various recall and estimation techniques to answer the stylized last week question. Of the 29 participants interviewed, 10 reported that they could recall directly how much sleep they got last week, either because they had a very structured schedule or had looked at the clock. The remainder used some other strategy to arrive at their answer. Nine participants reported using a rate retrieval strategy, where they estimated they slept about 7 hours per night. Participants who used a rate and adjustment or calculation strategy reported lower sleep estimates, around 5.5 hours of sleep. It is possible these participants adjusted their estimates downward too much due to calculation errors (Edgar, 2009).

Table 3. Frequency of recall or estimation strategies used to answer the stylized last week question

Strategy	**Frequency**
Recalled directly	10
Rate retrieval	9
Rate and adjustment	4
Calculation	4
Estimate/Guess	2

Reporting. We asked participants about possible social desirability concerns in reporting on sleep. For instance, we asked if they believed there is an appropriate number of hours that people should sleep in one night, and the minimum and maximum number of hours a person should sleep per night. Of the 29 participants, 21 indicated they believed there is an appropriate number of hours people should sleep in one night, while 8 indicated that it depends on the individual. On average, participants reported that 7.5 hours was the appropriate amount of sleep, with a range between 6.0 and 9.0 hours.

Of the 21 participants who indicated they believed there is an appropriate number of hours people should sleep in one night, all of them reported sleeping less than that amount. However, sleep estimates only deviated by an average of 25 minutes between participants' self-

reported appropriate sleep duration and diary-reported sleep duration. Deviations were higher for the stylized estimates, approximately 1.45 hours less than their self-reported appropriate sleep duration and their general and last week stylized-reported sleep duration.

Participants reported that oversleeping would be more embarrassing than undersleeping. Common examples of reasons for embarrassment at oversleeping included "looking lazy" or "being teased for sleeping too much." Participants were also asked whether they would be more embarrassed at over- versus under-sleeping if the survey was about employment or health. Of the 29 participants, 17 thought it would be embarrassing to admit sleeping too much in a survey about employment and jobs, and 13 thought it would be embarrassing to admit sleeping too little in a survey about health. Thus, survey context may differentially impact social desirability concerns. Sleeping too much in the context of a survey about employment may appear "lazy," whereas undersleeping in the context of a survey about health may appear as though a respondent does not get adequate sleep.

Summary. Through the cognitive interviews, we uncovered factors at each stage of the survey response process that may affect reports of sleep duration across diary and stylized measures and contribute to the sleep gap. We also found that, like the national surveys, participants reported getting more sleep in the diary than stylized measure. These insights allowed us to generate hypotheses about what factors might contribute to differences between diary and stylized sleep reports and the measurement error associated with each of them. For instance, perhaps providing respondents with a standardized definition of what counts as "sleep" would bring diary and stylized sleep estimates closer together. A survey about health versus jobs may cause people to report getting more or less sleep, respectively. However, the qualitative nature of the cognitive interviews limited the generalizability of the findings and the ability to make statistical comparisons. To determine whether these factors affect reports of sleep in diary and stylized measures, we conducted a larger scale, quantitative study.

1.5 Study 3: Quantitative Study

A large-scale online experiment collecting quantitative data was designed to make statistical comparisons of reported sleep duration across diary and stylized measures and test hypotheses generated from Studies 1 and 2. First, we wanted to determine whether we would replicate the sleep gap in our online sample. Based on the results of the cognitive interview study, we also wanted to determine whether providing a definition of sleep affects reported sleep duration. Finally, we wanted to assess context effects by comparing sleep duration estimates across participants who thought the survey was about jobs versus health.

Method. Participants were recruited using Amazon.com's Mechanical Turk (MTurk) platform, an online crowdsourcing website where research participants receive small incentives for completing surveys or other tasks (Buhrmester, Kwang, & Gosling, 2011). Although MTurk samples are not representative of the United States population, MTurk yields samples that are large and more demographically diverse than those obtained in other convenience samples, such as college students or local participants brought into cognitive laboratories (Casey, Chandler, Levine, Proctor, & Strolovitch, 2017; Stewart, Chandler, & Paolacci, 2017; Edgar, Murphy, & Keating, 2016). Participants were routed to a web survey, where they completed a modified version of the ATUS daily recall diary interview and answered a set of stylized questions about their activities (e.g., working, physical exercise, sleep). Again, participants completed the survey considering weekdays (both the diary day and previous 24-hour period were weekdays).

Participants and Design. A total of 1,233 participants living in the U.S. (54% female, with an average age of 36.34) complete the survey. A total of 62.1% were employed full time; 22.0% employed part-time; 9.4% were unemployed; 3.5% were students; and 3.0% were retired, and the average household size was 2.63. Demographics did not vary by condition ($ps > 0.08$). Participants were randomly assigned to one of three different survey framing conditions, in which they were told the survey was about health, jobs, or general time use. They were then randomly

assigned definitions of terms (including sleep[7]) or received no definitions. Finally, the order in which participants completed the diary and stylized questions in the survey was randomized. This yielded a 3x2x2 mixed-model design with 2 between-subjects factors (framing and definitions) and 1 within-subjects factor (question type). See Table 4 below for an illustration of the different survey variants.

Table 4. Experimental Design for the Quantitative Study

Between-Subjects Factors		**Within-Subjects Factors**
Survey Framing	**Activity definitions provided**	**Type of Measure (Order Randomized)**
Framing of "Survey about Jobs and Employment"	• Sleep definition • No definition	• Diary then Stylized • Stylized then Diary
Framing of "Survey about Health and Wellness"	• Sleep definition • No definition	• Diary then Stylized • Stylized then Diary
Framing of "Survey about Time Use" (control)	• Sleep definition • No definition	• Diary then Stylized • Stylized then Diary

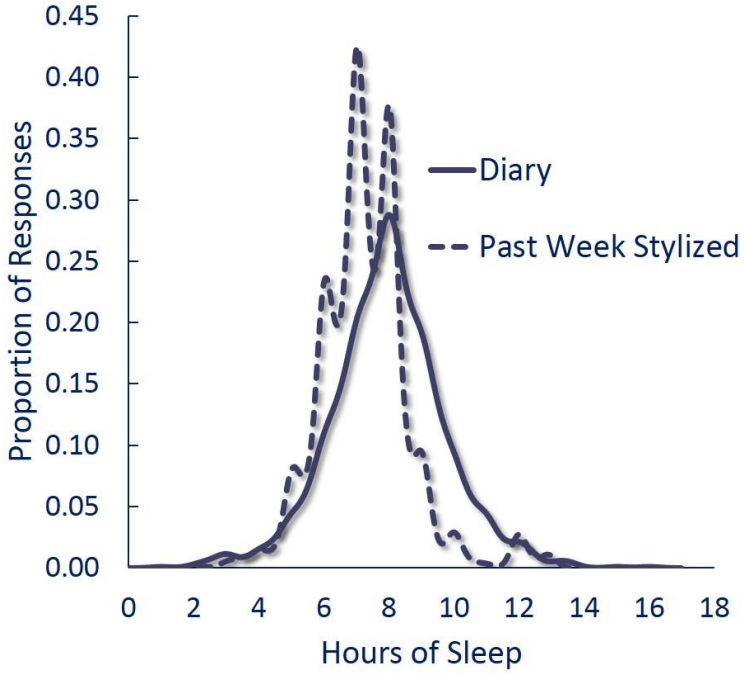
Participants completed an online version of the ATUS time diary in which they entered all of their activities from the prior 24-hour period (from 4 a.m. to 4 a.m.). They selected activities from a subset of the 12 most commonly reported activities in the ATUS (e.g., working, commuting) from a dropdown menu, and indicated the start and stop time of each activity, entering up to a maximum of 20 activities. They also answered a set of stylized questions about their activities throughout the previous week. Embedded within these questions were the same stylized sleep measures used in Study 2.

[7] The definition read, "By sleep, we mean the number of hours you actually spend sleeping. This may be different from the number of hours you spend in your bed, time you spend preparing to go to sleep, or resting with your eyes closed but not actually asleep. Please include any times you were sleeping during the day (or napping)." This was embedded with definitions of other common activities, such as work and exercise.

1.5.1 Quantitative Study Results

We calculated the total time participants reported sleeping in the prior 24-hour period to time diary estimate. We then compared participants' self-reported diary versus stylized hours of found a sleep gap where the diary measure (*M* = 7.95 hours; *SD* = 1.76) exceeded stylized reports (*M* = 7.27 hours; *SD* = 1.39) of sleep. As seen in Figure 1, we found the participants' stylized sleep reports peaked at rounded numbers, such as 6, and 7, and dropped quickly after 8 hours, indicating potential measurement error due to rounding.

Figure 1. Response distribution of sleep reports across diary and past week stylized questions.



To examine the effects and interactions of framing, definitions, and question order, we conducted a 3 (Framing type – jobs vs. health vs. time use) X 2 (Definitions – provided a definition vs. no definition) X 2 (Question Type Order – diary first vs. stylized first) mixed-model ANOVA, where the dependent variable was mean sleep duration as measured by the diary and past week stylized questions. Table 5 shows the results of this analysis.

Table 5. Results of Mixed-Model ANOVA on Sleep Duration by Question Type (Framing Type Definition), and Order (Dependent variable = Mean hours of sleep)

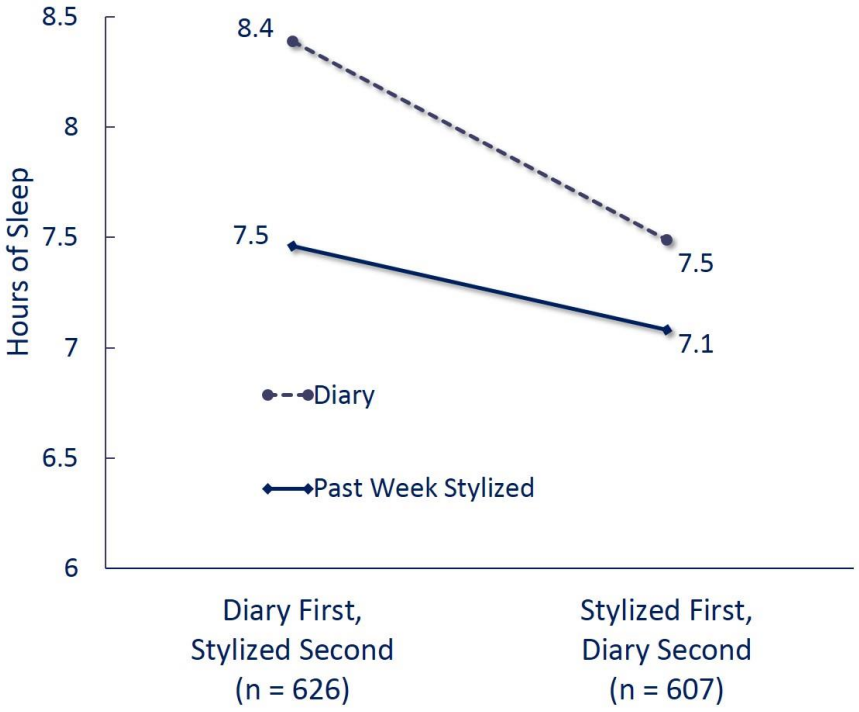
	df	**F Value**	**p-value**	**η^2_p**
Question Type	1	200.50	$p < 0.01$	0.14
Question Type X Definition	1	5.84	$p = 0.01$	0.01
Question Type X Order	1	29.49	$p < 0.01$	0.02
Question Type X Framing	2	1.55	$p = 0.21$	0.00
Question Type X Definition X Order	1	0.16	$p = 0.69$	0.00
Question Type X Definition X Framing	2	0.77	$p = 0.46$	0.00
Question Type X Order X Framing	2	0.26	$p = 0.77$	0.00
Question Type X Definition X Order X Framing	2	1.46	$p = 0.23$	0.00
Residuals	1221			

As seen in Table 5, we did not find a significant three-way interaction between definition, question order, and framing on sleep question type. We did find a significant interaction between question type and definition where participants who read the definitions reported more similar hours of sleep across the diary ($M = 7.87$, $SD = 1.74$) and stylized ($M = 7.33$, $SD = 1.52$) measures versus those who did not read definitions for the diary ($M = 8.00$, $SD = 1.77$) and stylized ($M = 7.22$, $SD = 1.26$) measures.

An interaction between question type and order was found, where participants who completed the diary first ($M = 8.39$ hours; $SD = 1.82$) reported more sleep in the time diary than participants who completed the stylized questions first ($M = 7.49$ hours; $SD = 1.57$) see Figure 2. In contrast, participants

who completed the stylized questions first (*M* = 7.08 hours; *SD* = 1.40) reported less sleep for the stylized question than those who completed the diary questions first (*M* = 7.47 hours; *SD* = 1.37).

Figure 2. Mean hours of sleep reported as a function of whether participants answered the diary vs. stylized questions first.



We also conducted a Fisher r-to-z transformation to test the magnitude of the correlation between the two measures. We found the correlation between the diary and stylized measures was significantly greater ($z = -4.38$, $p < .001$) when participants answered the stylized questions first ($r = 0.54$), versus completing the diary first, ($r = 0.34$). Thus, answering the stylized question first seemed to pull the diary and stylized sleep measures closer together. Finally, we found a main effect of question type where the diary sleep measure (*M* = 7.95; *SD* = 1.76) exceeded the stylized sleep measure (*M* = 7.28; *SD* = 1.40), replicating the sleep gap. No other significant main effects or interactions were found.

Summary. In the quantitative study, we found a sleep gap, where diary sleep estimates exceeded stylized sleep estimates. We also found that certain features of diary and stylized sleep questions (definitions, question order) may have a larger impact than others (survey context) on a web survey, as the framing of the survey (health, jobs, or time use) did not affect participants' reports of sleep. Context effects may have been minimized because the survey was self-administered and anonymous, which reduces social desirability concerns (Kreuter, Presser, & Tourangeau, 2008). Also consistent with the previous studies, providing a definition of sleep brought diary and stylized estimates slightly closer together, pointing to potential measurement error due to comprehension of the term "sleep."

We found that when stylized questions preceded diary questions, the gap between diary and stylized estimates was much smaller. This may indicate that participants alter their response strategies (e.g., recall or estimation) when answering both diary and stylized questions within the same survey. Although we did not anticipate this question order effect, stylized questions focus attention on a particular topic (in this case sleep), and this focus on sleep may have affected answers to subsequent questions (Schulz & Grunow, 2012). In contrast, when the time diary came first, the total amount of sleep reported was masked and embedded within reports of many other activities. This is consistent with the literature on order effects, showing that when specific questions (e.g., stylized questions) precede general, broader questions (e.g., diary questions), respondents anchor their answers to the more specific question that came first (Schwarz et al., 1991).

Our research so far has relied solely on self-reported sleep duration and could not assess whether diary or stylized sleep measures were more accurate. In our final study, we aimed to assess the accuracy of diary and stylized sleep measures for individuals by comparing them to sleep recorded by a sensor.

1.6 Study 4: Validation Study

The use of sensor data to conduct behavioral research has become more common over the past few years, providing an alternative, objective measure that is free from the measurement error associated with self-reports (Evenson, Goto, & Furberg, 2015; Wright, Brown, Collier, & Sandberg, 2017). As these technologies have emerged, researchers have become interested in whether they can be used to validate

survey questions (e.g., Downs, Van Hoomissen, Lafrenz, & Julka, 2014). We tracked respondents' activities (including sleep) to compare their self-reported sleep to objectively-measured sleep via a wearable device (the Fitbit Charge). We selected the Fitbit Charge because it automatically records sleep without any user action and does not display sleep data on its interface, so participants cannot view their sleep data. The Fitbit Charge has sensors that automatically measure the angle and movement of the device. The device then interprets these measurements as physical actions (e.g., walking or running). The device also measures the absence of movement, or detects only subtle movements, which is interpreted as sleep. Wearable devices are light and unobtrusive so they can be worn most of the time and during sleep. It has recently been suggested that the popularity of such devices may facilitate conducting larger-scale studies that compare self-reported sleep to objectively measured sleep (e.g., Miller et al., 2015). However, the validity and reliability of these devices varies, sometimes overestimating or underestimating different activities (Evenson et al., 2015). However, for sleep measurement, they are considered to be fairly reliable in healthy adults without sleep disorders (e.g., Kang et al., 2017; Lee et al., 2017; Cook, Prairie, & Plante, 2017).

Method. Participants were interviewed twice about one week apart. At Visit 1, they answered general questions about their typical routine and were instructed to wear the Fitbit Charge at all times over the next week, except while showering or bathing.[8] At Visit 2, participants completed the same abbreviated ATUS diary interview used in Study 2. We also asked participants a set of stylized sleep questions (the same ones used in Studies 2 and 3). Afterward, we compared the total sleep duration from the diary, stylized questions, and Fitbit measures, asking participants targeted probe questions to understand the differences (if any) between the measures. Amongst other data, the Fitbit records the number of hours slept per day and periods of wakefulness or restlessness during the night.

[8] When removing the device for bathing/showering respondents were asked not to clasp it shut so as to avoid the device registering sleep.

Participants. We recruited 44 participants in the Washington, DC metro area. Only participants who wore the Fitbit the day before the second visit were included in analyses to ensure sleep comparisons were possible across the diary, stylized questions, and Fitbit-recorded sleep durations.[9] A total of 35 participants (13 female with an average age of 44.58 years) complied with these instructions and were included in our final analyses.

1.6.1 Validation Study Results

Table 6 shows the mean sleep duration across each of the sleep measures. The Fitbit- estimate of sleep fell in between the diary and stylized questions estimates.

Table 6. Mean hours of sleep across diary, stylized, and Fitbit measures (N = 35).

Measure	Mean and SD Hours of Sleep
Diary	7.26 (1.79)
General Stylized	6.62 (1.14)
Last Week Stylized	6.56 (1.14)
Fitbit (prior 24 hours)	7.11 (1.64)
Fitbit (over past week)	6.88 (1.36)

We conducted a within-subjects ANOVA contrasting each of the sleep measures to assess whether participants' sleep duration estimates differed by sleep measure. We found the measures differed significantly, $F(4, 136) = 3.48$, $p = 0.10$, $\eta^2_p = 0.09$. Post-hoc pairwise comparisons showed that the diary and Fitbit-recorded sleep from the previous 24-hour period exceeded both stylized measures, ($ps < 0.05$). No other differences were found. We also assessed how well each measure agreed with one another using Intraclass Correlation Coefficients (ICC's).[10] Table 7 below shows the ICC's, confidence intervals, significance level, and agreement between each of the measures.

[9] Two participants lost the device during the week, six could not return for their second interview, and one participant's device fell off during the night before their scheduled interview.

[10] ICC's were calculated because this statistic provides a measure of how well related variables (e.g., sleep measures from the same participant) that measure the same construct agree with one another (see Kang et al., 2017). Results using a Spearman correlation coefficient were similar to those obtained using the ICC's.

Table 7. Intraclass Correlation Coefficients (ICC's) and Confidence Intervals (CI's) across Sleep Duration Measures

Measures	ICC and CI	*p*-value	Agreement
Diary and Stylized (weekdays)	0.49 (0.19 - 0.71)	$p = 0.002$	Fair
Diary and Fitbit-recorded sleep (prior 24 hours)	0.76 (0.60 - 0.88)	$p < 0.001$	Excellent
Stylized and Fitbit-recorded sleep (average over week)	0.40 (0.08 - 0.64)	$p = 0.01$	Fair
Stylized and Fitbit-recorded sleep (average over weekdays)	0.62 (0.35 - 0.79)	$p < 0.001$	Good
Stylized and Fitbit-recorded sleep (average over weekend)	0.30 (0.05 - 0.58)	$p > 0.05$, n.s.	Poor

The diary and stylized measures had fair agreement with one another, consistent with prior research (e.g., Schulz & Grunow, 2012). The diary and Fitbit-recorded sleep from the previous 24-hour period yielded the best agreement among all of the measures, falling in the excellent agreement range, consistent with literature showing that diary measures may be a more reliable measure of time use (Juster et al., 2003; Kan & Pudney, 2008). Overall, the stylized and Fitbit-recorded sleep over the week had fair agreement, with good agreement on weekdays and poor agreement on weekends. This is consistent with our findings from the cognitive interviews, suggesting that people may be better at estimating their sleep and wake times on weekdays, when they tend to follow a more structured schedule, versus weekends where schedules are less structured and estimating sleep duration may be more difficult. Similar to Study 3, we again observed rounding in the stylized measure, with participants providing responses of five, six, or seven hours of sleep.

During respondent debriefing, we found that in some instances the Fitbit-recorded data aided participants' recall of their wake and sleep times, (e.g., recalling they hit the snooze button and got some extra sleep, or woke up a little earlier than normal on that day). Participants could generally recall waking up during the night once or twice, but the Fitbit tended to show many awakenings during the night that participants could not recall. Thus, the Fitbit may have its own set of measurement error where the absence of movement does not always correspond to sleep and movement does not always correspond to being awake (Wright et al., 2017). For example, the Fitbit may have overestimated the amount of wakefulness experienced during the night (e.g.,

recording tossing and turning as time awake), underestimating the total sleep duration for each night. In other cases, the Fitbit may have overestimated sleep, recording period of time lying still watching television or reading as naps.

Summary. We found evidence that sleep duration recorded via sensor data may fall somewhere in between diary and stylized sleep estimates. Diary measures tended to agree more with the sensor data, consistent with prior research showing that diary measures may be more reliable and valid than stylized measures. However, sensor data are also prone to measurement and user error, and the Fitbit-recorded sleep may not always accurately reflect actual sleep.

1.7 General Discussion

This research investigated the gap between diary and stylized sleep measures and potential sources of measurement error associated with them. We used different questionnaire evaluation methods (behavior coding, cognitive interviews, quantitative research, and a validation study) to address our research questions. In Table 8 below, we summarize the main findings from each method, the sources of measurement error each uncovered, and the pros and cons associated with each method.

We found a sleep gap across Studies 2-4, where diary sleep measures led participants to report more sleep than stylized measures. Each method revealed sources of measurement error that may have caused diary measures to exceed stylized measures, helping to explain reasons for the sleep gap.

Study 1 used behavior coding to identify issues in the ATUS interviews that may lead to measurement error. We found that interviewer and respondent interactions surrounding sleep are often complex. Respondents had difficulty recalling or estimating their sleep and wake times. Interviewers often used leading questions and unscripted probes that may encourage respondents to believe they should define sleep as a continuous episode, potentially inflating ATUS sleep estimates. One strength of behavior coding is the use of actual production interviews to identify potential issues and the ability to investigate and code numerous features of the diary interview, including questions, probes, and answers, as well as unscripted conversation that may contribute to error in the ATUS. A limitation is that it did not provide

direct insight into how respondents arrived at their answers to these questions, and we had no direct comparison to stylized questions. This led us to conduct cognitive interviews.

Table 8. Summary of the main findings, sources of measurement, and pros and cons of each method

	Sleep gap observed?	Comprehension	Recall	Judgment	Reporting	Sources of Measurement Error	Pros of Method	Cons of Method
Behavior coding (Study 1)	N/A	N/A	Respondents used alarms or TV programs to help recall wake and sleep times	More conversational turns and qualifications occurred in reporting sleep time than wake times	N/A	-Interactions surrounding sleep times are complex, may be imprecise -Interviewers use unscripted, leading probes that may affect ATUS sleep estimates	-Use of production interviews -Code numerous features of ATUS interview -Identify problematic concepts, tasks, and practices	-Little insight into respondents' cognitive processes -No comparison to stylized sleep questions
Cognitive interviews (Study 2)	Yes	Broad definitions of sleep were associated with reporting more sleep than narrow definitions of sleep	Respondents used alarms or TV programs to help recall wake and sleep times	Respondents used rate retrieval, rate and adjustment, calculation, and guessing for stylized questions	Survey context (employment vs. health) may affect self-reports of sleep	-Recall and estimation bias may be present in reporting sleep times -Survey context may push sleep estimates up or down based on social desirability	-Rich understanding of cognitive processes surrounding sleep questions -Hypothesis generation	-Small, non-representative sample -Cannot make statistical inferences and comparisons
Quantitative study (Study 3)	Yes	Providing a definition of sleep narrowed the sleep gap	N/A	Rounding observed in stylized sleep estimates	No survey framing or context effect observed (employment vs. health) on self-reports of sleep	Definitions of sleep and question order affected self-reports of sleep, but not survey framing	-Collected large amount of data in short timeframe -Large sample allowed for statistical comparisons	-Non-probability sample, cannot make generalizations to the U.S. population
Validation study (Study 4)	Yes	N/A	Sensor data aided recall for sleep and wake times	Rounding observed in stylized sleep estimates	N/A	-Sensor data agreed more with diary sleep estimates -Wearable devices have their own set of measurement and user error to consider	-Objective sleep measure does not rely on self-report data	-Sensor-recorded data also prone to measurement error -User error with wearable device

Study 2 involved conducting cognitive interviews with the aim of understanding how respondents report on sleep at each stage of the response process. Building on our knowledge from the behavior coding research, we found that definitions of sleep, recall of sleep times, and social desirability biases were areas where measurement error is likely to occur. The cognitive interviews provided a rich understanding of these issues, and allowed us to generate hypotheses about what factors might contribute to differences between diary and stylized sleep reports. One downside was the small, geographically limited sample that could not be used to make statistical comparisons. This led us to conduct a larger-scale quantitative study.

In Study 3, we designed a quantitative experimental study to compare participants' diary and stylized sleep estimates, where we also found a sleep gap in which diary measures exceeded stylized measures of sleep duration. Drawing on the results of the previous two studies, we found that how respondents define sleep affected their answers for the diary and stylized last week measures, but not the general stylized measure. It may be that people rely on the typical amount of sleep they get overall (e.g., 7 hours) when reporting in general versus considering activities that took place over the week, an area for future investigation. We also found that when stylized questions preceded diary questions, the sleep gap narrowed, indicating that participants may have anchored their answers to the stylized estimate. Context effects were less apparent in the quantitative study – perhaps due to mode – being an online, anonymous survey rather than an interviewer-administered survey (e.g., Kreuter et al., 2008).

One benefit of using online crowdsourcing panels such as Mechanical Turk is that it enabled us to collect a large amount of data in a short timeframe (Edgar et al., 2016). We were able to obtain a larger, more geographically diverse sample from participants around the country than would be possible to obtain in traditional laboratory studies, such as cognitive interviews (Casey et al., 2017; Stewart et al., 2017). Although MTurk samples are not representative of the general population, they are useful for experimental purposes and as a research tool since we were interested in assessing internal validity rather than representativeness of any particular population. As such, crowdsourced panels are not a replacement for probability samples, and the

results should be interpreted with caution. Also, these findings could not tell us whether diary or stylized sleep estimates were more accurate, which led us to our validation study.

Finally, in Study 4, we conducted a validation study that compared sleep duration across diary, stylized, and sensor data. We found evidence that an objective measure of sleep may fall somewhere in between diary and stylized measures, but the sensor data agreed more with diary than stylized self-reports. Viewing the sensor data also helped participants recall their sleep and wake times in some instances; however, it was not without its own set of measurement and user error. Depending on the researcher's goals, such devices could be a useful tool to assess sources of question measurement error for surveys where participants are asked to recall their activities or time use.

1.8 Implications and Future Directions

These results have broad implications for researchers interested in measuring time use. Researchers should be aware that diary and stylized questions might yield different results and understand the sources of measurement error associated with both measures. Future research should explore additional reasons for the gap between diary and stylized measures beyond just the response process. This might include sampling, context effects, interviewer effects in the administration of both diary and stylized questions, data collection procedures, and how the survey organization defines and calculates time spent on activities. Researchers might also explore other activities that show a gap in diary and stylized measures, such as work or exercise. As wearable devices improve, researchers may want to capitalize on these new technologies to investigate potential sources of survey measurement error. In future research, we also recommend using a multi-method approach (e.g., D'Ardenne & Collins, in press), as each method can capture unique sources of measurement error and can build off the preceding results and insights. We believe this approach will be highly useful to researchers designing, evaluating, testing, or validating survey questions.

References

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., &

Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology, 161*(4), 389-398.

Biddle, J. E., & Hamermesh, D. S. (1990). Sleep and the allocation of time. *Journal of Political Economy, 98*, 922-943.

Bonke, J. (2005). Paid work and unpaid work: Diary information versus questionnaire information. *Social Indicators Research*, *70*(3), 349-368.

Brenner, P. S. (2011). Identity importance and the overreporting of religious service attendance: Multiple imputation of religious attendance using the American Time Use Study and the General Social Survey. *Journal for the Scientific Study of Religion, 50*(1), 103-115.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Canfield, B., Miller, K., Beatty, P., Whitaker, K., Calvillo, A., Wilson, B. A. (2003). Adult questions on the Health Interview Survey - Results of cognitive testing. Internal NCHS report.

Casey, L., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. (2017). Demographic characteristics of a large sample of us workers. Retrieved from https://osf.io/preprints/psyarxiv/8352x/

Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, 6(4), 339-366.

Cook, J. D., Prairie, M. L., & Plante, D. T. (2017). Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorders*, *217*, 299-305.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, 71(4), 623-634.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

D'Ardenne, J. and Collins, D. (in Press) Combining multiple question evaluation methods – what does it mean when the data appear to conflict? In Beatty, P., Collins, D., Kaye, L., Padilla, J.L., Willis, G., and Wilmot, A. (Eds.). *Advances in questionnaire design, development, evaluation, and testing.* Hoboken, NJ: Wiley.

Denton, S., Edgar, J., Fricker, S., and Phipps, P. (2012). Exploring conversational interviewing in the American Time Use Study: Behavior coding study report. Internal BLS report.

Downs, A., Van Hoomissen, J., Lafrenz, A., & Julka, D. L. (2014). Accelerometer-measured versus self-reported physical activity in college students: Implications for research and practice. Journal of American College Health, 62(3), 204-212.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. *Survey measurement and process quality*, 287-310.

Edgar, J. (2009). What does "Usual" usually mean? Paper presented at the American Association for Public Opinion Research.

Edgar, J., Murphy, J., & Keating, M. (2016). Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions. *SAGE Open*, *6*(4), 2158244016671770.

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1), 1.

Ford, E. S., Cunningham, T. J., & Croft, J. B. (2015). Trends in self-reported sleep duration among US adults from 1985 to 2012. *Sleep*, *38*(5), 829-832.

Hale, L. (2005). Who has time to sleep? *Journal of Public Health*, *27*, 205-211.

Hurst, M. (2008). Who gets any sleep these days? Sleep patterns of Canadians, *Canadian Social Trends*, *85*. Statistics Canada Catalogue no. 11-008-XWE. Retrieved from: http://www.statcan.gc.ca/pub/11-008-x/2008001/article/10553-eng.htm

Juster, F. T., Ono, H., & Stafford, F. P. (2003). An assessment of alternative measures of time use. *Sociological Methodology*, *33*(1), 19-54.

Kan, M. Y. (2008). Measuring housework participation: the gap between "stylised" questionnaire estimates and diary-based estimates. *Social Indicators Research*, *86*(3), 381-400.

Kan, M. Y., & Pudney, S. (2008). Measurement error in stylized and diary data on time use. *Sociological Methodology*, *38*(1), 101-132.

Kang, S. G., Kang, J. M., Ko, K. P., Park, S. C., Mariani, S., & Weng, J. (2017). Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of Psychosomatic Research*, *97*, 38-44.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: the effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

Lee, H. A., Lee, H. J., Moon, J. H., Lee, T., Kim, M. G., & Kim, L. (2017). Comparison of wearable activity tracker with actigraphy for sleep evaluation and circadian rest-activity rhythm measurement in healthy young adults. *Psychiatry investigation*, *14*(2), 179-185.

Lin, K. H. (2012). Revisiting the gap between stylized and diary estimates of market work time. *Social science research*, *41*(2), 380-391.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Miller, C. B., Gordon, C. J., Toubia, L., Bartlett, D. J., Grunstein, R. R., D'Rozario, A. L., & Marshall, N. S. (2015). Agreement between simple questions about sleep duration and sleep diaries in a large online survey. *Sleep Health*, *1*(2), 133-137.

Phipps, P. A., & Vernon, M. K. (2009). Twenty-four hours: An overview of the recall diary method and data quality in the American Time Use Survey, pp. 109-128. In Robert F. Belli, Frank P. Stafford,

and Duane F. Alwin (Eds.) *Calendar and Time Diary Methods in Life Course Research*, Thousand Oaks, CA: Sage.

Schulz, F., & Grunow, D. (2012). Comparing diary and survey estimates on time use. *European Sociological Review, 28(5),* 622-632.

Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public opinion quarterly, 55(1),* 3-23.

Silva, G. E., Goodwin, J. L., Sherrill, D. L., Arnold, J. L., Bootzin, R. R., Smith, T., Walsleben, J. A., Baldwin, C. M., & Quan, S. F. (2007). Relationship between reported and measured sleep times: The sleep heart health study (SHHS). *Journal of Clinical Sleep Medicine, 3*, 622- 630.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences*.

Tourangeau, R. (1984). Cognitive sciences and survey methods. *Cognitive aspects of survey methodology: Building a bridge between disciplines*, 73-100.

Van der Zouwen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behavior codes and question–answer sequences: a diagnostic approach. *Methods for testing and evaluating survey questionnaires*, 109-130.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Wright, S. P., Brown, T. S. H., Collier, S. R., & Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *312(3),* 358-367.