Hitczenko, Marcin

Working Paper

Improved estimation of poisson rate distributions through a multi-mode survey design

Working Paper, No. 2021-10

Provided in Cooperation with:
Federal Reserve Bank of Atlanta

Suggested Citation: Hitczenko, Marcin (2021) : Improved estimation of poisson rate distributions through a multi-mode survey design, Working Paper, No. 2021-10, Federal Reserve Bank of Atlanta, Atlanta, GA, https://doi.org/10.29338/wp2021-10

This Version is available at:
https://hdl.handle.net/10419/244313

FEDERAL RESERVE BANK *of* ATLANTA **WORKING PAPER SERIES**

Improved Estimation of Poisson Rate Distributions through a Multi-Mode Survey Design

Marcin Hitczenko

Working Paper 2021-10
February 2021

Abstract: Researchers interested in studying the frequency of events or behaviors among a population must rely on count data provided by sampled individuals. Often, this involves a decision between live event counting, such as a behavioral diary, and recalled aggregate counts. Diaries are generally more accurate, but their greater cost and respondent burden generally yield less data. The choice of survey mode, therefore, involves a potential tradeoff between bias and variance of estimators. I use a case study comparing inferences about payment instrument use based on different survey designs to illustrate this dilemma. I then use a simulation study to show how and under what conditions a hybrid survey design can improve efficiency of estimation, in terms of mean-squared error. Overall, this work suggests that such a hybrid design can have considerable benefits as long as there is nontrivial overlap in the diary and recall samples.

JEL classification: C15, C81, C83

Key words: recall surveys, diaries, bias, mean-squared error, multi-level models

https://doi.org/10.29338/wp2021-10

Please address questions regarding content to Marcin Hitczenko, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309, marcin.hitczenko@atl.frb.org.

1 Introduction

Much research in the social sciences involves the study of rates: how frequently people act in certain ways or experience certain events. Indeed, individual count data is found in datasets relevant to a variety of fields, including economic consumption (BHPS; CES; SCA; PSID), health (NHIS), media (BMCS), and crime (BRFSS), among others (see the reference list for the full names of these datasets). In particular, the example used in this work relates to the study of payment instrument use among consumers.

A researcher often has a choice of how to collect such count data from sampled individuals. In this work, we juxtapose two modes of data collection: "live" data collection, in which events are recorded as they occur, and recall surveys, in which respondents provide a retrospective event count for a pre-specified period of time. Live data collection can take many different forms, but perhaps the most common is the behavioral diary, in which respondents track daily events as they happen. From this point on, we focus primarily on diary data, though the ideas in this work apply to other forms of live data collection.

The appeal of the recall survey directly relates to its logistical advantages. As opposed to a recall query, diaries are generally more difficult to implement and demand a greater respondent burden, leading to a higher cost per respondent. Beyond that, diary fatigue, in which respondents' motivation wanes as the length of the observation period increases, suggests limiting the length of diary measurement periods to maintain suitable data quality (Ahmed et al. 2010; Jonker and Kosse 2009; Silberstein and Scott 1991; Schmidt 2011). As an example, most consumer payment diaries organized by Central Banks last from a day (Netherlands) to three days (United States and Canada) to a week (Germany, France, Austria, and Australia). On the other hand, months (used in the CES) and years (used in the PSID and the SCA) have all been used in recall surveys. The difference in cost can be such that, within a fixed budget, a recall survey collects data from more individuals and for longer observation periods than a diary.

Unfortunately, recalled count data are notoriously subject to error. Both omission and telescoping, wrongly counting events that occurred outside the period in question, have been documented in past studies (Bound et al. 2001; Groves 1989; Neter and Waksberg 1964). In fact, research suggests that the dependability of recall is governed by a complex cognitive process

(see Rockwood (2015) for an overview). In general, accuracy of recall is linked to saliency, a somewhat nebulous concept relating to the frequency, regularity, and impact of the event in question. Social desirability has also been shown to lead to over-reporting of seemingly commendable activities, such as exercising, and under-reporting of negatively-perceived behavior, such as drug use (Shephard 2003; Tourangeau and Yan 2007).

Of course, diary data is not immune to inaccuracies. Much like longitudinal studies, diaries are subject to attrition and the aforementioned diary fatigue, which can introduce nonignorable response bias when the loss of data is linked to the behavior of interest (Groves et al. 2001; Thomas et al. 2016). Some multi-day diaries, such as the Consumer Expenditure Survey, observe significant data entry at the end of the observation period, thus nudging it in the direction of recall and jeopardizing quality (Silberstein and Scott 1991; Crossley and Winter 2014). Finally, it has been hypothesized that the act of recording one's behavior itself may result in unusual behavior on the part of the individual, although there has been no conclusive evidence to verify this hypothesis (Kemsley and Nicholson 1960; McKenzie 1983).

The attributes of each survey mode have implications on the quality of inference, introducing potential tradeoffs between bias and variance. In this paper, we consider the possible benefits of a hybrid design that combines diary and recall data. To do so, we assume that a diary likely represents a higher standard of data than a recall survey, which we reduce to an assumption that diary counts are accurate and recalled counts are potentially inaccurate and systematically biased. This general notion is supported by research on topics as diverse as reporting food consumption (Brzozowski et al. 2017), hospital visits (Clarke et al. 2008), exercise (Nusser et al. 2012), household chores (Marini and Shelton 1993), and job-related accidents (Andersen and Mikkelsen 2008). Moreover, the quality of diary data is likely to generally improve with the increased implementation of new technology that makes mobile tracking and data entry easier and more reliable(Anderson et al. 2016; Chatzitheochari et al. 2018; Greaves et al. 2015; Siemieniako 2017). As a result, we believe the ideas in this work have the potential to benefit research in many fields.

We begin by specifying the research problem and developing a general framework of analysis in Section 2. Section 3 uses a case study to show how inference based on different modes can lead to different results. The methodology of assimilating the two data modes is developed in

Section 4, and a simulation study is used to determine the extent of the potential gains and how they can be practically factored into survey design. Finally, we discuss the general findings and the implications for survey design in Section 6.

2 Framework

Although the example in this work relates to research on the frequency of payment instrument use, the ideas are relevant to any study of how often individuals experience certain events. No matter the discipline, the unifying framework is a population of individuals, indexed by the subscript i, with associated rates, μ_i. Each rate, μ_i, defines the expected number of events experienced by individual i for a chosen reference period. It is assumed that $\mu_i \sim F(\boldsymbol{\theta})$ for some family of distributions, $F(\cdot)$, and a set of parameters, $\boldsymbol{\theta}$. The researcher is interested in estimating $\boldsymbol{\theta}$ for a particular $F(\cdot)$.

Information about μ_i comes from collected count data corresponding to a measurement period of length ℓ. A common assumption is that the reported counts follow a Poisson distribution with parameter defined in part by μ_i. The assumed distributions for the observed count data and the rates combine to form a hierarchical model that can be used to estimate $\boldsymbol{\theta}$.

2.1 Diary vs. Recall

One measure of the information collected in a dataset of counts is the total length of time observed, generally a number of days. There are two dimensions to this; the number of respondents in the sample, N, and the number of days of observation for each individual ℓ, so that a total of $N\ell$ days are observed. In diaries, ℓ corresponds to the number of days of tracking, while in a recall survey, it represents the length of the recall period.

As a simple, illustrative example, consider the case of a homogeneous Poisson point process with daily rate μ_i, so that individual i's reported count for ℓ days is $C_i \sim \text{Poisson}(\ell\mu_i)$. Then, for a given sample of size N, a natural estimate of the mean population rate is $\frac{1}{N\ell}\sum_{i=1}^{N} C_i$. If sampling of individuals is appropriately representative, the bias of this estimate is zero, and the mean-squared error reduces to $N^{-1}\left(\text{Var}(\mu_i) + \frac{\text{E}(\mu_i)}{\ell}\right)$.

As expected, an increase in the sample size results in a lower mean-squared error, as does a lengthening of the recall period, though the latter does so with a non-zero lower bound. Determining N and ℓ to minimize mean-squared error within a fixed budget depends not only on the first two moments of the rate distribution, but on the relative costs of increasing the sample size versus extending the measurement period. A data product that combines a greater sample size and a longer measurement period, such as a recall survey, is clearly preferable if bias is not a concern. However, the nature of the recall bias is generally unknown, making the diary a safer, though less precise, option, and thus presenting the researcher with a dilemma regarding survey mode.

2.2 Evaluating Survey Design

A necessary component of this study is evaluating survey designs, which we do through the average quality of inference associated with data generated via said survey design. More formally, let $\mathcal{S}$ represent a set of specifications and instructions for generating a dataset, including sample size, recruitment methodology, questionnaire design, and any other aspects that affect the nature of the data or how it is analyzed. We let $\mathsf{data}_k(\mathcal{S})$ represent a random dataset drawn according to the specifications defined by $\mathcal{S}$, with the subscript k indexing unique, independently drawn datasets. Adopting a Bayesian paradigm, the data are used to generate a posterior estimate for $\boldsymbol{\theta}$, which we label $\boldsymbol{\theta}_k(\mathcal{S})$. The model and methodology used to estimate the posterior distribution are incorporated into $\mathcal{S}$.

A simple measure of how well the posterior distribution estimates any parameter $\theta \in \boldsymbol{\theta}$ is the mean-squared error:

$$\mathrm{MSE}(\theta_k(\mathcal{S})) \quad = \quad \mathrm{E}\left[\theta_k(\mathcal{S}) - \theta\right]^2.$$

In the simulations found in this paper, we consider a special case where $\mathsf{data}_k(\mathcal{S}) \subset \mathsf{data}_k(\mathcal{S}')$. Then, the ratio of mean-squared errors,

$$\Phi_k(\theta \mid \mathcal{S} \to \mathcal{S}') = \frac{\mathrm{MSE}(\theta_k(\mathcal{S}))}{\mathrm{MSE}(\theta_k(\mathcal{S}'))}, \tag{1}$$

measures the benefit in efficiency of the new data in $\mathcal{S}'$. The average of (1) with respect to the distribution of possible datasets under sampling schemes $\mathcal{S}$ and $\mathcal{S}'$,

$$\Phi(\theta \mid \mathcal{S} \to \mathcal{S}') \;=\; \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \Phi_k(\theta \mid \mathcal{S} \to \mathcal{S}').$$

quantifies the added value of the additional information in $\mathcal{S}'$ relative to that in $\mathcal{S}$. The closer that $\Phi(\theta \mid \mathcal{S}, \mathcal{S}')$ is to zero, the larger fraction of information about the parameters is featured in the added data. Additionally, an identity such as $\Phi(\theta \mid \mathcal{S} \to \mathcal{S}') < \Phi(\theta \mid \mathcal{S} \to \mathcal{S}'')$ suggests that the survey design $\mathcal{S}'$ is preferable to $\mathcal{S}''$.

3 Case Study: Frequency of Payment Instrument Use

Before turning to the simulation, we consider a case study using data from different survey designs to infer the frequency of payment instrument use of cash, credit cards, debit cards, and checks among likely adopters of each payment instrument. The restriction to likely adopters is made to avoid more complex models that must accommodate bimodal distributions due to reported zeros by nonadopters.

For each payment instrument, there are five datasets. One is extracted from the 2012 Diary of Consumer Payment Choice (DCPC), a three-day diary of all payment activity. The other four are recall-based data specific to each payment instrument for recall periods of a day, week, month, and a year, from RAND survey "Well Being 199", which we dub the 2011-2012 Recall Survey. Both datasets were collected from members of RAND's American Life Panel (ALP), the details of which can be found at `www.RAND.org/ALP`. Each dataset is a reasonable representation of what a researcher studying such questions might have available, and there is no prior reason to think they should not be used to make inferences. Despite this, we find that estimates based on all five data constructs yield significantly different results.

As the case study is primarily an illustrative example of survey mode effects, the exposition is deliberately concise and is organized as follows. Section 3.1 introduces the general model of payment behavior that informs all later analysis. A brief summary of the data source and the likelihoods used in diary-based and recall-based estimation are given in Section 3.2 and Section

3.3, respectively. Finally, we provide details of parameter estimation and discuss the results in Section 3.4.

3.1 Model

The most basic unit of observation is the number of payments made by individual i on any given day t, which we model as

$$C_{it} \sim \text{Poisson}(\mu_{it}), \tag{2}$$

where μ_{it} can be decomposed into an individual-specific rate and an effect corresponding to the day of the week of day t. Thus, we let

$$\texttt{dow}(t) = \begin{cases} 1 & t \text{ is a Sunday} \\ 2 & t \text{ is a Monday} \\ \vdots & \vdots \\ 7 & t \text{ is a Saturday}, \end{cases}$$

so that

$$\log(\mu_{it}) \;\; = \;\; \log(\mu_i) + \lambda_{\texttt{dow}(t)}. \tag{3}$$

We further enforce that $\sum_{d=1}^{7} \mathrm{e}^{\lambda_d} = 1$, so that μ_i represents a weekly rate. Figure 1 shows the daily sample averages and spreads for the 35 days in 2012 for which DCPC data are collected. Except for a jump in check use on the first and, to a lesser extent, the last day of the month, Figure 1 suggests that a large part of temporal variation in daily behavior can be attributed to day-of-week effects. For check use, this is predominantly defined by less use on Saturday and Sunday. Cash, credit cards, and debit cards, on the other hand, show greater overall use on Friday and Saturday.

In practice, temporal patterns are almost certain to be more complicated and heterogeneous. Nevertheless, the weekly cycle seems to represent a decent approximation to a complex reality

for some period of time surrounding the dates of observation. Even this approximation may not apply to different time periods due to seasonal trends that have systematic impacts on individuals' payment behavior.

The weekly rates, μ_i, are assumed to have the following distribution

$$\begin{aligned} \log(\mu_i) &= \mu + \alpha_1 \texttt{age}(i) + \alpha_2 \texttt{edu}(i) + \alpha_3 \texttt{inc}(i) + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \sigma^2), \end{aligned} \tag{4}$$

where $\texttt{age}, \texttt{edu}, \texttt{inc}$ describe the age, education level, and household income of individual i. All three are treated as numerical variables, and the education and household income levels associated with each numeric value are shown in Appendix A.

3.2 Estimation Based on Diary Data

3.2.1 Diary Data: Source

The 2012 DCPC invited $2{,}505$ individuals from the ALP to record various aspects of their payment behavior for three consecutive days randomly assigned between September 30^{th} and November 2^{nd}. Over the three days, respondents track and record details of all of their personal financial transactions, including payments. Diary respondents are asked to enter information about their daily transactions in an online module at the end of each day of participation. To help keep track of transactions, respondents are mailed and encouraged to use two paper memory aids and a pouch in which they can keep receipts. Almost 90 percent of respondents enter data within 24 hours of the diary day for which they are reporting, and over 95 percent do so within 3 days. Even if recall is used, the diary respondent benefits from prior knowledge that transactions are to be reported as well as a relatively short gap between the transaction and it's recording.

3.2.2 Diary Data: Pre-Processing

The number of purchases made with each payment instrument on each day can be extracted from the 2012 DCPC data. Because general purchases and bills are reported in separate modules, it is possible to enter a bill payment once in each. Thus, we only count once entries in the bill

and general purchase modules that share the same payment instrument, amount of payment, and merchant for a given individual and day of reporting. The result for each individual is a triplet of daily number of purchases, $D_i = \{D_{it_{i1}}, D_{it_{i2}}, D_{it_{i3}}\}$, corresponding to the three days of participation, t_{ij}, $j = 1, 2, 3$.

Likely payment instrument adopters are identified with the help of the 2012 Survey of Consumer Payment Choice (SCPC), a second payments survey with high overlap with the DCPC, which directly asks about ownership and use within the past year of various payment instruments. In 2012, $2{,}348$ of the diarists also participated in the SCPC. Respondents are classified as likely adopters if they report at least one use of the payment instrument in the diary or if they claimed adoption in the SCPC. The final number of likely adopters within the 2012 DCPC are 2,467 for cash. 1,857 for credit card, 2,075 for debit card, and 2,146 for check.

3.2.3 Diary Data: Likelihood

Mirroring the model developed in Section 3.1, we assume $D_{it} \sim \text{Poisson}(\mu_{it})$ with the mean μ_{it} decomposed as in (3). The data likelihood function assumes not only independence across respondents but also a conditional independence between an individual's daily counts given $\{\mu_{it}\}$:

$$\text{Prob}\left(\{D_i\} \mid \{\mu_{it}\}\right) \;=\; \prod_i \prod_{j=1}^{3} \text{Prob}\left(D_{it_{ij}} \mid \mu_{it}\right). \tag{5}$$

3.3 Estimation Based on Recall Data

3.3.1 Recall Data: Source

The 2011-2012 Payment Recall Survey is an effort led by RAND to study the quality of recall via five online surveys administered every three months between July 2011 and September 2012 to a starting field of 3,516 ALP panelists. In each survey, respondents recall the number and total dollar value of payments made with each of the four payment instruments for four different recall periods: day ($\ell = 1$), week ($\ell = 7$), month ($\ell = 31$), and year ($\ell = 365$). Across surveys, the framework of recall would vary, alternatively asking for a specific period of time and a "typical"

period of time. A more detailed description of the full data can be found in Angrisani et al. (2014), but we focus on the subset of 1,285 respondents who provided the number of uses of each payment instrument for specific recall periods corresponding from July to September of 2012.

Figure 2, which shows the dates of the recall survey and the first day of the diary for the 715 individuals who were featured in both, offers a representative view of participation dates. Although all surveys are emailed on the 15^{th} of each month, respondents can take the survey whenever they want. The specific recall periods are assigned at the commencement of the survey, so the reported values are relative to the day on which the survey was taken rather than to the day the survey link was emailed. Daily recall is asked for a randomly selected day in the week prior to the survey, an effective way to ensure uniform observations across the days of the week. Longer recall periods directly precede the recall survey. Recall is done for each payment instrument sequentially, with the order of the instruments chosen at random. In addition, for each payment instrument the order of the daily, weekly, and monthly periods is randomized, with the yearly period always coming last.

3.3.2 Recall Data: Pre-Processing

In the case of recall data, likely adopters are defined as anyone who claims to be an adopter in the 2012 SCPC (977 recall survey respondents participated in the 2012 SCPC), anyone who made a payment in the 2012 DCPC, or, for the 289 who did not participate in either, reported making payments for at least one recall period in the recall survey. A necessary part of using the recall data for estimation is addressing highly unlikely response numbers in the right tail that are likely to affect parameter estimates. In this analysis, we adopt the approach of limiting estimation to those responses below some threshold. Specifically, let μ_{max} be the supposed maximum weekly rate, so that the number of payments in a period of ℓ days is approximated by $\text{Poisson}\left(\frac{\ell}{7}\mu_{max}\right)$. For each recall period, ℓ, we take the 95^{th} quantile and discard all responses over this threshold. Table 2 shows the thresholds, the number of likely adopters, and the number of observations above the threshold for each payment instrument and each recall period.

3.3.3 Recall Data: Likelihoods

Letting $\{R_{i\ell}\}$ be the set of recall responses for period length ℓ, we assume independence across individuals with $R_{i\ell} \mid \mu^*_{i\ell} \sim \text{Poisson}(\mu^*_{i\ell})$, where the asterisk identifies parameters associated with recall. By estimating $\boldsymbol{\theta}$ from each recall period separately, we ignore any dependence between reported numbers for different recall periods, potentially caused by having a reported count for one period anchor those for subsequent periods (Means et al. 1989; Sudman et al. 1996). However, such dependencies, if anything, suggest there should be more consistency in the estimated rates across recall periods than if anchoring were avoided.

3.3.3.1 Daily Recall Because it seems plausible that the survey lag in daily recall affects the quality of recall (Sudman and Bradburn 1973), we incorporate its potential effect into the daily count model given in (3). We define $\texttt{lag}(i) = 0, \ldots, 6$ as the number of days between the recall survey and the day in question, with $\texttt{lag}(i) = 0$ indicating that the day of the recall survey directly the follows the day for which counts are requested. The reported count for assigned day s_{i1} is assumed to have mean defined by

$$\log(\mu^*_{i1}) \;=\; \log(\mu_i) + \lambda_{\texttt{dow}(s_{i1})} + \gamma \texttt{lag}(i). \tag{6}$$

The model in (6) assumes that reported recall corresponds to the true behavior when the survey lag is zero and that the effect of the survey lag effect is monotonic. While more complicated dynamics may be more realistic, it is counterintuitive that they are non-monotonic or that greater accuracy comes from a longer survey lag.

3.3.3.2 Weekly/Monthly/Yearly Recall For longer recall periods the survey lag is zero, and we define the Poisson mean by

$$\log(\mu^*_{i\ell}) \;=\; \log(\mu_i) + \log\left(\frac{\ell}{7}\right). \tag{7}$$

In the case of monthly and yearly recall, the form in (7) is an approximation of the true rate. Let $s_{i\ell}$ be the start of the recall period and $e_{i\ell} = s_{i\ell} + \ell$ the end of the recall period. Then, assuming

conditional independence across daily counts within an individual given individual rates, the mean number of payments for individual i's recall period is given by the sum of the relevant daily means:

$$\mu^*_{i\ell} \;=\; \sum_{t=s_{i\ell}}^{e_{i\ell}} \mu_{it} = \mu_i \sum_{d=1}^{7} \mathrm{e}^{\lambda_d} \mathtt{k}_{i\ell}(d),$$

where $\mathtt{k}_{i\ell}(d)$ represents the number of times day-of-week d appears in individual i's ℓ-day recall period. For weekly recall, $\mathtt{k}_{i7}(d) = 1$ for all d, which combines with the restriction $\sum_{d=1}^{7} \mathrm{e}^{\lambda_d} = 1$ to yield $\mu^*_{i7} = \mu_i$, as implied by (7). In any 31 day period, $\mathtt{k}_t(i, \ell)$ will be 4 for four consecutive days and 5 for three consecutive days, instead of the 4.42 of each implied by the approximation. Based on the estimates of daily effects from the diary data, the percent difference between the smoothed approximation and the true rate is no more than 0.9 percent for cash, credit, and debit, and no more than 1.4 percent for check, where day-of-week effects are more pronounced. In any 365-day period, $\mathtt{k}_t(i, \ell)$ will be 52 for six days and 53 for one day, meaning the maximum percent difference between approximated mean and the true mean is less than 0.1 percent for all payment instruments. Conceptually, the simplification in (7) mirrors the cognitive recall process for longer periods, in which the episodic recall and enumeration used for shorter periods (Bradburn et al. 1987; Strube 1987) is replaced by rate-based approximation (Blair and Burton 1986; Eisenhower et al. 1991; Menon 1994).

3.4 Parameter Estimation

3.4.1 Estimation Details

To ease interpretation of parameters, all three demographic variables are centered and standardized by dividing by twice the standard deviation of the observed values in the sample of all diarists, as advocated by Gelman (2008). The priors taken for the primary parameters of interest,

$\boldsymbol{\theta} = \{\mu, \alpha_1, \alpha_2, \alpha_3, \sigma\}$ are

$$\begin{aligned} \mu &\sim \text{Normal}(0, 2) \\ \alpha_s &\sim \text{Normal}(0, 1), s = 1, \ldots, 3 \\ \sigma &\sim \text{Exp}(1). \end{aligned} \tag{8}$$

For analysis that involves accounting for day-of-week effects, namely the diary data and the daily recall, parameter estimation is simpler without a restriction on the sum of day-of-week effects. In that case, the weekly rate is not represented by μ_i, but must be calculated by summing over the daily rates: $\sum_{d=1}^{7} \mu_i e^{\lambda_d}$. For the day-of-week effects, we assume a prior of

$$\lambda_d \sim \text{Normal}(0, 2), d = 1, \ldots, 7. \tag{9}$$

The daily recall model, given in (6), also estimates the survey lag effect, for which we use the prior

$$\gamma \sim \text{Normal}(0, 2).$$

All models are fit with `R-STAN`, using 4 chains and 3,000 iterations for each chain with a burn-in period of 1,500 iterations of the MCMC algorithm. To estimate posterior distributions for each parameter, we thin by drawing every 10^{th} iteration, thus ending up with 600 posterior draws. Diagnostics of the MCMC suggest proper performance. The Gelman-Rubin convergence statistic, $\hat{R}$, is near 1 for all parameters, suggesting convergence of the chains (Gelman and Rubin 1992). In addition, trace plots for each chain suggest good mixing and stationarity, and posterior means are very similar to those when the model is estimated with `glmer`.

3.4.2 Results

In comparing the estimated dynamics based on different data sources, we focus on the demographic means. These are characterized by the slopes in (5), α_s, as well as the base mean, which defines the expected value for an individual with standardized demographic values of zero. For

a Log-Normal distribution, this base mean is given by $\mathrm{E}\left[\mu_i \mid \texttt{demo}_i\right] = \mu + \frac{\sigma^2}{2}$ with respect to parameters defined in (5). Figure 3 shows means and 95 percent credible intervals for each of the four parameters based on the five data sources.

The diary results show some interesting results regarding how demographics affect payment instrument use. Credit cards are used more frequently with increasing age, though it seems that higher income and education levels are the greater driving force behind use. Conversely, check use is primarily driven by age, with older individuals using checks more frequently. Debit card use decreases with age, and there is generally more homogeneity across social strata. Finally, the use of cash is generally steady across demographic groups.

In comparing the recall-based estimates to those based on the diary, perhaps the most obvious finding is that the base mean is poorly estimated by all four recall surveys. Estimates based on daily recall are especially poor, even when accounting for the effect of survey lag, which has a minor impact: posterior means range from -0.04 to 0.02 and posterior standard deviations ranging from 0.03 (cash) to 0.06 (check). Except in the case of cash, the base means are consistently overestimated in the recall surveys. This phenomenon is consistent with findings in other fields that show that recalled data often over-estimate diary-based estimates (Ahmed et al. 2010; Clarke et al. 2008; Nusser et al. 2012).

On the other hand, the three longer recall periods do reasonably well at estimating the marginal demographic effects. Of the 12 slopes estimated, the credible intervals based on recall data overlap with the diary interval all but once each for monthly and yearly recall and in all cases for weekly recall. For daily recall, there is overlap in only eight cases. Subsequently, the posterior mean based on recall falls within the diary-based interval six to eight times for the longer recall periods and only three times for daily recall.

Comparisons between the diary and recall data are not perfect. Some fraction of the discrepancies in the findings can be attributed to seasonal differences between observation periods or the methodology used to define likely adopters or clean the data, but these seem unlikely to fully explain the observed inconsistencies. Thus, if one assumes the diary as accurate, it follows that recall data yields fundamentally incorrect inferences about population dynamics, most notably regarding the baseline number of weekly payments. Moreover, different recall periods yield the most accurate results for different payment instruments. Although we are unaware

of other analyses comparing diary and recalled payments data specifically, the observed inconsistency across recall periods is to be expected based on more general research on consumption (Ahmed et al. 2010; Deaton and Grosh 2000; Hurd and Rohwedder 2009; NSSO Expert Group on Sampling Errors 2003).

4 Alternative Survey Design: Simulation Study

In this section, we use a simulation framework to study the potential benefits of a survey design in which diary data, assumed to be unbiased, is supplemented with possibly erroneous recall survey data. The basic methodology is to directly model and estimate the discrepancy between diary and recall rates within the process of estimating $\boldsymbol{\theta}$. The hope is that, although potentially inaccurate, the recall data contains enough information about true rates to outweigh the value of replaced diaries. After conducting a simple simulation, we discuss how our results can be applied to improve efficiency of surveys in practice.

4.1 Simulation

We consider a simulation framework similar to that of the case study. Thus, weekly means are defined by the following identities:

$$\begin{aligned} \log(\mu_i) &= \mu + \alpha\, X_i + \epsilon_i \\ X_i &\sim \text{Normal}(0, 1) \\ \epsilon_i &\sim \text{Normal}(0, \sigma^2). \end{aligned} \tag{10}$$

A hypothetical researcher is interested in estimating $\boldsymbol{\theta} = \{\mu, \alpha, \sigma\}$. One option is to field a 3-day diary, in which the reported numbers follow the same distribution as the truth:

$$D_i \mid \mu_i \sim \text{Poisson}\left(\frac{3}{7}\mu_i\right). \tag{11}$$

Alternatively, the researcher can rely on recall for the past month (31 days) which is potentially associated with recall error:

$$\begin{aligned} R_i \mid \mu_i^* &\sim \text{Poisson}\left(\frac{31}{7}\mu_i^*\right) \\ \mu_i^* \mid \mu_i, \mu_e, \sigma_e &\sim \text{LogNormal}(\mu_i + \mu_e, \sigma_e^2). \end{aligned} \tag{12}$$

The nuisance parameters are μ_e, σ_e, with the former identifying a systematic bias, a tendency to either over- or underestimate, and the latter defining how correlated the recall rate is to the real rate, μ_i. Specifically, when σ_e is small, μ_i^* is close to $\mu_i + \mu_e$, with larger values of σ_e allowing greater deviations from $\mu_i + \mu_e$.

We consider simulations for eight different scenarios defined by two different models of the truth and four different models for recall error, the details of which are specified in left-most table in Table 3. With respect to true behavior, Model 1 is roughly based on the parameters corresponding to cash use, while Model 2 is based on those of check use. The average weekly rate for the former is about 3.1 with a standard deviation of 1.73, and over a three-day diary period, fewer than 1 percent of respondents will not have made a purchase. Model 2, on the other hand, has an average weekly rate of 0.66 and a standard deviation of 0.99, and we expect almost one-third of responses to have no payments in any three-day diary period.

The four types of recall are defined by degree of bias and variance and are shown in the right-most table in Table 3. A useful measure in assessing the quality of recall is the ratio of $\text{Var}[\mu_i] = \sigma^2 + \alpha^2$ to $\text{Var}[\mu_i \mid \mu_i^*] = \sigma^2 + \alpha^2 + \sigma_e^2$. A ratio close to zero indicates that the the variance in recall error makes it virtually impossible to decipher the true rate and makes the added recall data less valuable. In the low bias recalls, the ratios are 0.83 and 0.96 for Model 1 and Model 2, respectively, while high bias yields values of 0.12 and 0.41 respectively.

4.2 Analysis

There are many possible survey designs, but we focus on one in which $N = 1{,}000$ diaries are supplemented with $N = 1{,}000$ recall surveys. The only flexibility in the survey design is the degree of overlap between the two samples, or how many respondents provide dairy and recall

information. We characterize this by the parameter $p \in [0, 1]$. If $p = 0$, no respondents take both surveys, and a total of $2{,}000$ individuals need to be recruited. On the other extreme, if $p = 1$ only $1{,}000$ individuals are recruited and each takes both the recall survey and the diary. For each of the 8 sets of parameters, $\{\boldsymbol{\theta}, \mu_e, \sigma_e\}$, we consider $p = 0, 0.25, 0.5, 0.75, 1$. Within each of the 40 parameter configurations, we run 60 independent simulations, with iteration k proceeding as follows.

Step 1 Let $\mathcal{I}_k^d(p)$ represent a sample of $N = 1{,}000$ respondents chosen at random, and let

$$\text{data}_k^d(p) = \{D_i \mid i \in \mathcal{I}_k^d(p)\}.$$

represent the corresponding diary data.

Step 2 Use the hierarchical model specified by (11) and (11) along with the priors in (8) to generate a posterior distribution for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_k^d(p) = \boldsymbol{\theta}_k \mid \text{data}_k^d(p).$$

Step 3 For each $\theta \in \boldsymbol{\theta}$, we let θ_{kj}^d represent the j^{th} draw from the posterior. Based on 500 draws from the posterior, we estimate the mean-squared error:

$$\text{MSE}_k^d(\theta \mid p) \quad = \quad \frac{1}{500} \sum_{j=1}^{500} (\theta_{kj}^d(p) - \theta)^2.$$

Step 4 Let $\mathcal{I}_k^r(p)$ represent a set of $N = 1{,}000$ individuals chosen to take the recall survey such that the specifications implied by p. First, a random set of $p \times 1{,}000$ of individuals is taken from $\mathcal{I}_k^d(p)$ and then an additional $(1 - p) \times 1{,}000$ respondents are chosen at random for the recall survey. Let

$$\mathcal{I}_k^d(p) \bigcup \mathcal{I}_k^r(p)$$

will have exactly $1000p$ respondents selected at random. Let

$$\text{data}_k^h(p) = \{D_i \mid i \in \mathcal{I}_k^d(p)\} \bigcup \{R_i \mid i \in \mathcal{I}_k^r(p)\}.$$

represent the corresponding hybrid dataset.

Step 5 Use the hierarchical model specified by (11)-(12), along with priors in (8) and $\mu_e \sim \text{Normal}(0, 2)$ and $\sigma_e \sim \text{Exp}(1)$ to generate a posterior distribution estimate of $\boldsymbol{\theta}$, which we represent with:

$$\boldsymbol{\theta}_k^h(p) = \boldsymbol{\theta}_k \mid \text{data}_k^h(p).$$

Step 6 For each $\theta \in \boldsymbol{\theta}$, we let θ_{kj}^h represent the j^{th} draw from the posterior. Based on 500 draws from the posterior, we estimate the mean-squared error:

$$\text{MSE}_k^h(\theta \mid p) \quad = \quad \frac{1}{500} \sum_{j=1}^{500} (\theta_{kj}^h(p) - \theta)^2.$$

Step 7 Calculate the estimated ratio in parameter efficiency for each $\theta \in \boldsymbol{\theta}$:

$$\Phi_k(\theta \mid p) \quad = \quad \frac{\text{MSE}_k^h(p)}{\text{MSE}_k^d(p)}. \tag{13}$$

4.3 Results

The roots of the individual $\Phi_k(\theta \mid p)$ as well as the mean $\hat{\Phi}(\theta \mid p) = \frac{1}{60} \sum_{k=1}^{60} \hat{\Phi}_k(\theta \mid p)$ are shown in Figure 4. There are a few distinct phenomena that are generally delineated according to the variance of the recall error and whether or not there was overlap in the diary and recall samples. When there is overlap, or $p > 0$, the additional data improves average efficiency, though there is a clear dichotomy between cases with low variance error and those with high variance. When the recall is more closely correlated with the truth, the improvement is significant, with $\hat{\Phi}(\theta \mid p)$ averaging around 0.76 for all non-zero values of p. The percent of cases in which $\Phi_k(\theta \mid p) < 1$

ranges from about 65 to 95 percent. By contrast, when the recall variance is high, the average value of $\hat{\Phi}(\theta \mid p)$ is 0.95, and the percent of cases in which $\Phi_k(\theta \mid p) < 1$ is as low as 50 percent, when estimating μ and σ under Model 1, and peaks around 80 percent, when estimating α in Model 2. Essentially, when recall quality is bad, the additional data does not provide much information about true behavior and thus the parameters of interest. While the case where $p = 0.5$ tended to have the greatest efficiency improvements, the differences between cases where $p > 0$ were not significant.

An interesting phenomenon occurs when there is no overlap in the recall and diary samples. Then, the values of $\hat{\Phi}(\theta \mid 0)$ are generally above one and the added diary data decreased efficiency more often than not, even when variance of recall is low, meaning the added recall data actually made inference systematically worse. As seen in Figure 4, this is particularly egregious for estimating α. To further study the loss in efficiency of estimating α when $p = 0$, Figure 5 shows how the added recall data affect the posterior means and standard errors of the posterior estimates in the case where recall has low bias and low variance (top panel) and high bias and variance (bottom panel). Again, in cases where there is some overlap, $p > 0$, the hybrid-based estimates remain unbiased. However, the posterior variances reduces much more noticeably when recall variance is low than when it is high. On the other hand, when there is no overlap, the estimates of α are biased downwards, with the bias worst for the low variance recall error (bias is around -0.15 rather than -0.05). Intuitively, this might occur because, without overlap, the data are consistent with models in which recall rates are characterized by a simple translation of the true rates without additional intra-person variation, so that σ_e is small. Under such models, the intra-person noise has the effect of diffusing the correlation between X_i and the observed counts, thus shrinking α toward zero. In cases where there is overlap, it is much easier to correctly identify the intra-person variation, σ_e.

5 Improvement in Surveys

Up to this point, all analysis has ignored the costs of various survey designs, an essential factor in determining allocation of resources. In considering the practical implications of our simulation results, we again consider a researcher with a binary choice: collect M' diaries or collect M

diaries and M recall surveys. We call the two survey designs $\mathcal{S}_d$ and $\mathcal{S}_h$, respectively. The number of additional diaries sacrificed under the hybrid design depends on the relative cost of additional diaries and recall surveys. Letting c represent the ratio of the two and assuming that cost is proportional to the number of respondents, $M' = (1 + c)M$ corresponds to the same total cost under both designs.

The framework developed in this paper naturally allows a comparison of the two survey designs by determining the benefits of each relative to a base design, $\mathcal{S}_0$, in which only M diaries are collected. Under the notation of this paper, the comparison of interest is the relative efficiency improvements of each design, which can be measured by

$$\text{Relative Efficiency}(c) = \frac{\Phi(\theta \mid \mathcal{S}_0 \to \mathcal{S}_h)}{\Phi(\theta \mid \mathcal{S}_0 \to \mathcal{S}_d)}.$$

Based on theory outlined in Section 2, and validated by simulations with $N = 500$ and $N = 2{,}000$, $\Phi(\theta \mid \mathcal{S}_0 \to \mathcal{S}_1) = \frac{1}{1+c}$. We consider $\Phi(\theta \mid \mathcal{S}_0 \to \mathcal{S}_2) = 0.75, 0.85,$ and 0.95 based on values observed in the simulations. Figure 6 shows the relative efficiencies of the two survey designs as a function of c, so that the lines correspond to each survey design having the same total cost. Values less than 1 suggest the hybrid design is more efficient within the corresponding cost structure. Naturally, as the relative cost, c, increases the recall surveys have less worth, and it takes a greater saving to make the hybrid approach worth it. However, for a reasonable efficiency of 0.85, recall surveys can be as much as one-fifth the cost of diaries to have the hybrid design yield greater efficiency.

We note that the case where $p = 1$, which showed potentially significant efficiency gains in the simulations, corresponds to adding a recall question to a diary effort. In our experience, doing so is often associated with virtually no additional cost, as the extended burden is not great and no new respondents need to be recruited. Thus, it seems decidedly valuable to link a short recall survey before or after the diary period.

6 Discussion

Overall, our simulation study suggests that if the quality of recall is reasonable and the cost of recall surveys is not too great, supplementing diary observations with recall surveys, as long as there is non-trivial overlap in the samples, provides for more efficient estimators of population dynamics than an all-diary design. In such cases, the sheer amount of information contained in the additional surveys outweighs that in the replaced diaries. Of course, the simulations in this paper represent a relatively narrow range of possible survey designs and more varied analyses would be informative in understanding how to best allocate resources.

Consolidating diary and recall data requires assumptions about the nature of the recall error. While we believe the model introduced in this paper is adequate for many settings, extensions that further recognize subsets of the sample for which recall is of high quality can generate greater efficiency gains. For example, Battistin and Padula (2015) suggest that the levels of discrepancy depend on observable demographics such as income and education, in which case the bias and variance associated with recall can be made to differ according to demographic strata. A second possible development, based on the idea that lower-frequency events are more salient and, thus, better recalled, links individual recall error rates, now defined by μ_{ei} and σ_{ei}, to μ_i through functional form, perhaps $\sigma_{ei} = a\mu_i^b$.

The cost analysis developed in this paper is a simplified one intended to demonstrate the potential benefits of the hybrid design. In practice, researchers can use a similar approach to determine the number of diaries and surveys as well as degree of sample overlap that minimizes mean-squared error for a given budget, as long as the cost of each survey design can be calculated and the relative efficiency of a hybrid design can be approximated. An even more sophisticated approach might be an adaptive survey design, in which results are analyzed as data comes in. Thus, the relative value of the surveys is actively evaluated, and, based on this, the number of additional surveys and diaries that should be administered in the future is determined. If recall is proving to have no value, all future resources can be used for diaries before the entire budget is exhausted. Alternatively, unbiased recall surveys suggests using them exclusively.

A Appendix A

Value	Education Level	Household Income
1	Less than 1st grade	Less than $5,000
2	1st,2nd,3rd, or 4th grade	$5,000-$7,499
3	5th or 6th grade	$7,500-$9,999
4	7th or 8th grade	$10,000-$12,499
5	9th grade	$12,500-$14,999
6	10th grade	$15,000-$19,999
7	11th grade	$20,000-$24,9499
8	12th grade (no diploma)	$25,000-$29,999
9	High school graduate or GED	$30,000-$34,999
10	Some college, but no degree	$35,000-$39,999
11	Associate degree in occupational/vocational program	$40,000-$49,999
12	Associate degree in academic program	$50,000-$59,999
13	Bachelor's degree	$60,000-$74,999
14	Master's degree	$75,000-$99,999
15	Profession school degree	$100,000-$124,999
16	Doctorate degree	$125,000-$199,999
17		$200,000 or more

Table 1: Numeric values and their corresponding education levels and household incomes.

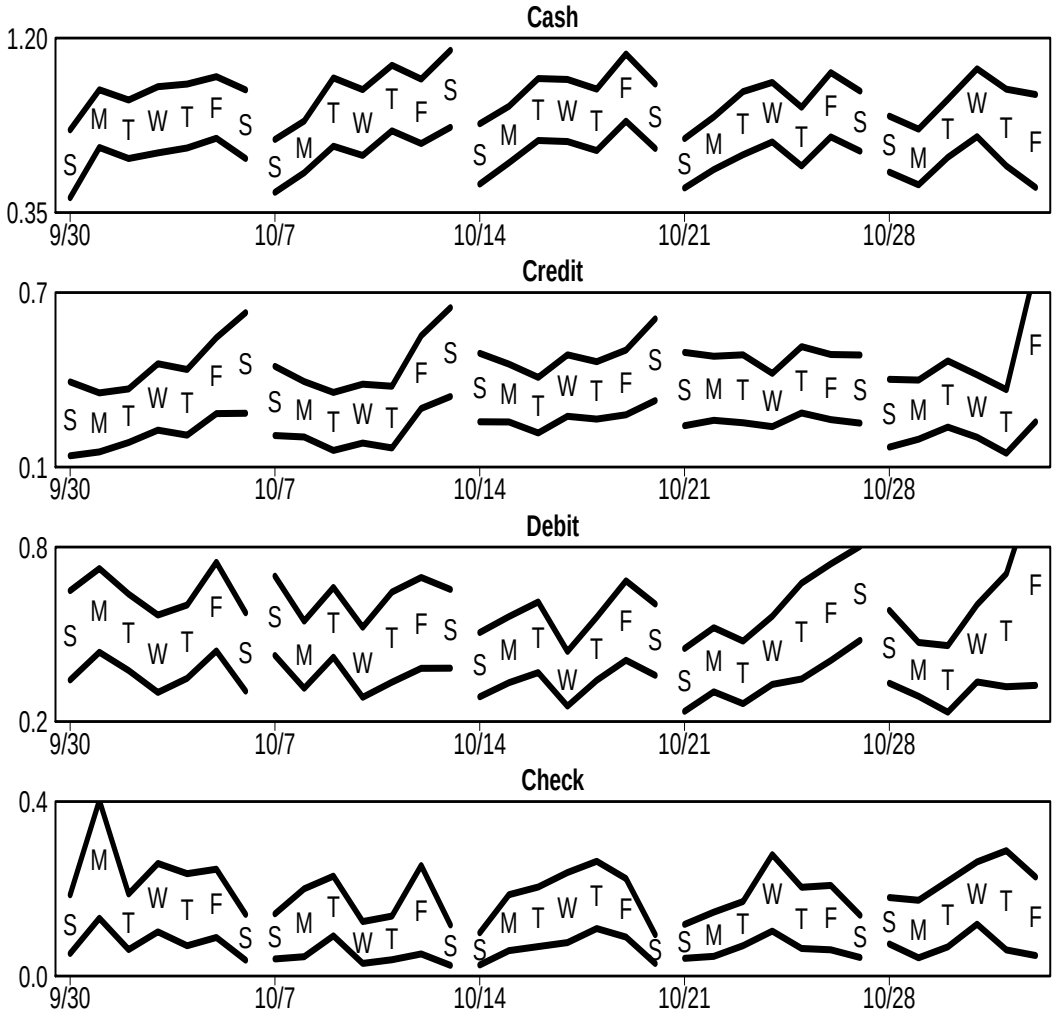

Figure 1: Daily averages and averages ± 2 standard deviations in the 2012 DCPC.

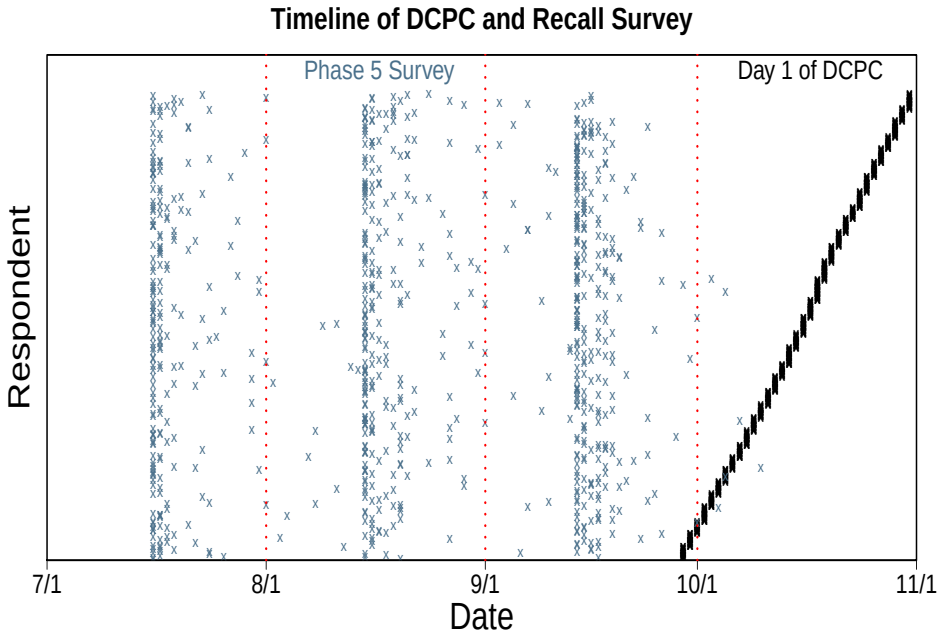

Figure 2: A temporal distribution of the recall survey and the DCPC for 715 individuals who participated in both.

$\mu_{max} = 75$					
Payment Instrument	# of Adopters	# Over Threshold			
		$\ell = 1$	$\ell = 7$	$\ell = 31$	$\ell = 365$
Cash	1240	38	36	26	26
Credit	919	9	14	18	14
Debit	925	16	26	19	16

$\mu_{max} = 50$					
Payment Instrument	# of Adopters	# Over Threshold			
		$\ell = 1$	$\ell = 7$	$\ell = 31$	$\ell = 365$
Check	990	3	8	11	20

Table 2: Number of recall responses above threshold removed from analysis and the total number of observations for likely adopters. When $\mu_{max} = 75$ the thresholds are 16, 90, 329, and 4,003 for daily, weekly, monthly, and yearly recall. When $\mu_{max} = 50$, the corresponding thresholds are 12, 62, 224, and 2,684.

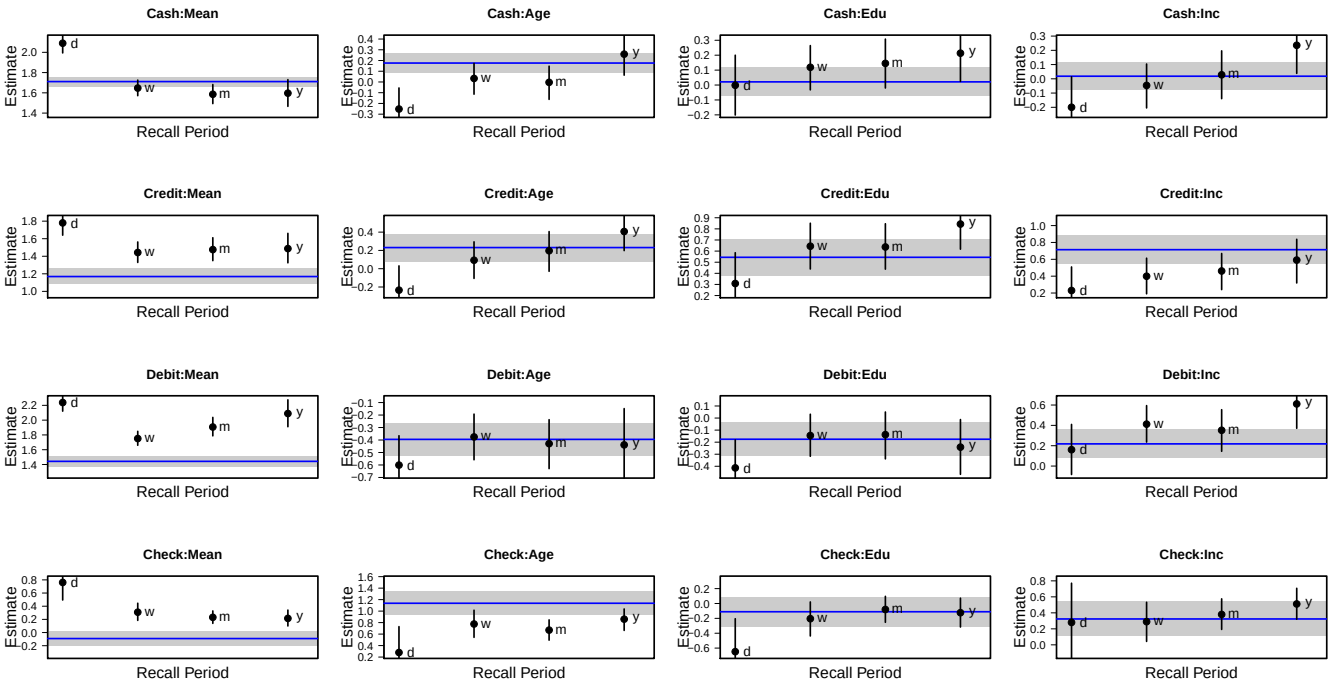

Figure 3: Parameter estimates and 95 percent intervals based on diary (gray bars), daily ("d"), weekly ("w"), monthly ("m"), and yearly ("y") recall.

True Behavior ($\boldsymbol{\theta}$)			
	μ	α	σ
Model 1	1	.25	.5
Model 2	-1	.75	1

Recall Error		
	μ_e	σ
Unbiased/Low Variance	0	0.25
Biased/High Variance	.5	1.5
Unbiased/High Variance	0	1.5
Biased/Low Variance	.5	0.25

Table 3: Parameters used in simulations: two defining true distribution of weekly means and four defining recall error.

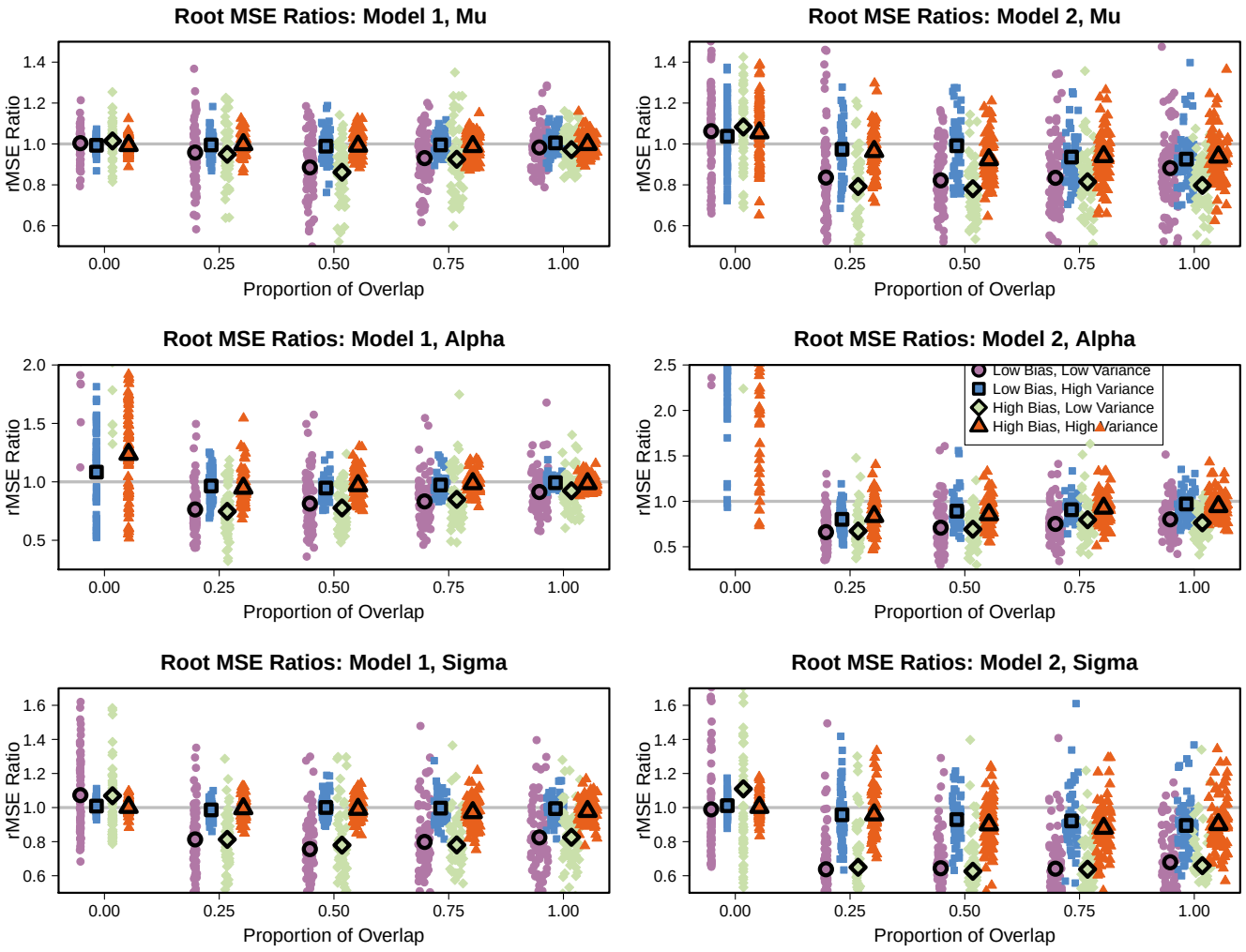

Figure 4: Observed values of $[\Phi_k(\theta \mid p)]^{1/2}$ and averages, $\left[\hat{\Phi}(\theta \mid p)\right]^{1/2}$ for different values of p, and different models of true behavior and recall error as defined in Table 3

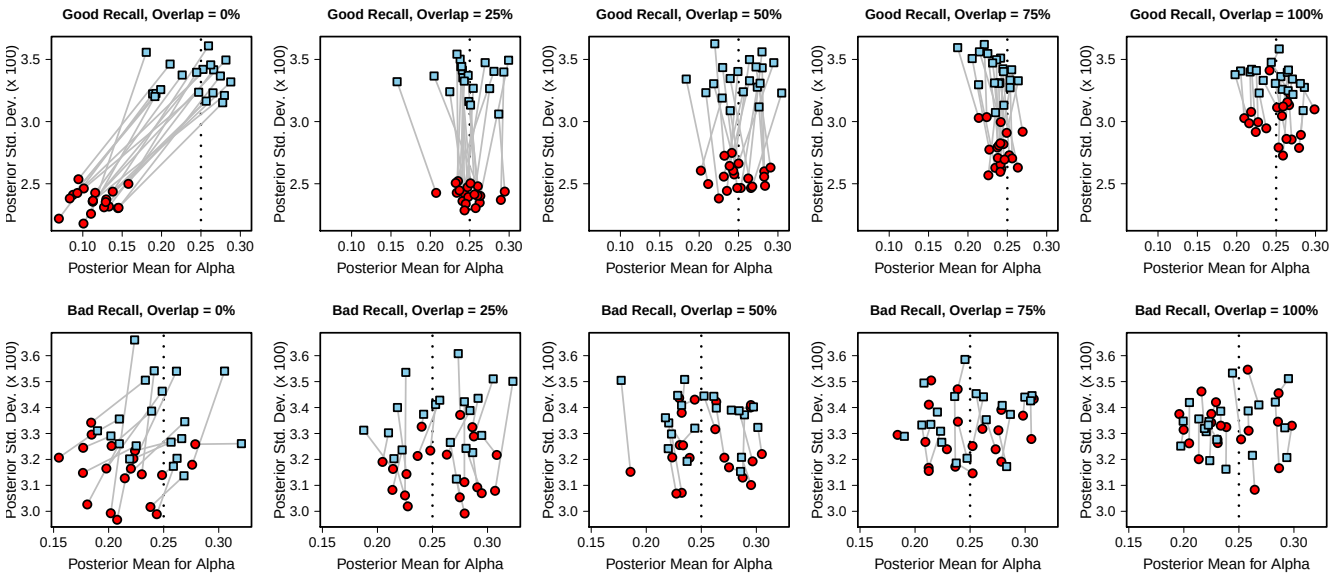

Figure 5: Posterior means and standard deviations of $\boldsymbol{\theta}$ based on diary data only (blue square) and hybrid data (red circle) for different degrees of overlap, p, for Model 1 and two different recall errors. "Good Recall" refers to the low bias and low variance recall error, and "Bad Recall" refers to the high bias and high variance recall error as defined in Table 3.

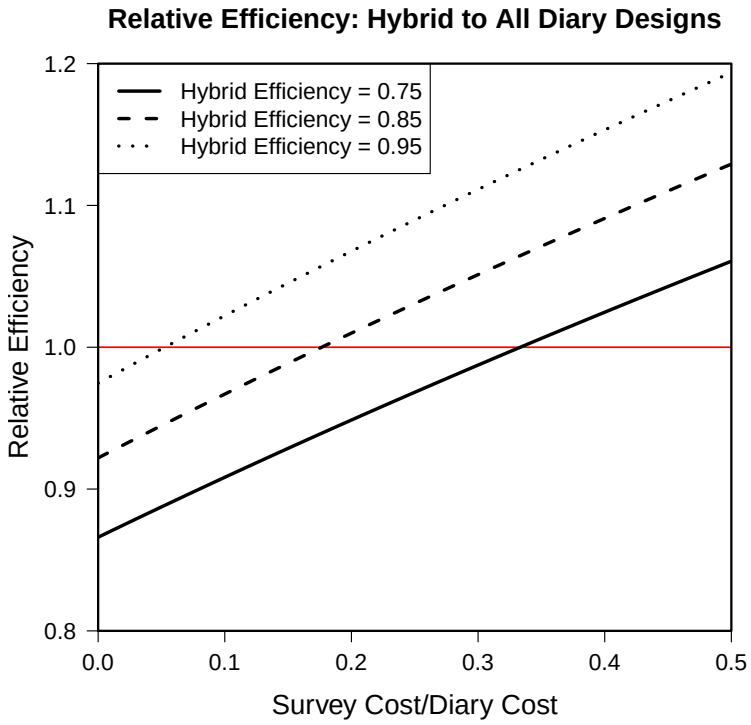

Figure 6: The relative efficiency of using a hybrid design to one in which only diaries are used based on the relative cost of the recall survey to the diary.

References

Ahmed, Naeem, Matthew Brzozowski, and Thoms F. Crossley. 2010. "Measurement Errors in Recall Food Consumption Data." *Institute for Fiscal Studies Working Papers* .

Andersen, L.P. and K.L. Mikkelsen. 2008. "Recall of Occupational Injuries: A Comparison of Questionnaire and Diary Data." *Safety Science* 46 (2):255–260.

Anderson, Kevin, Oksana Burford, and Lynne Emmerton. 2016. "Mobile Health Apps to Facilitate Self-Care: A Qualitative Study of User Experiences." *PLoS ONE* 11(5).

Angrisani, Marco, Arie Kapteyn, and Scott Schuh. 2014. "Measuring Household Spending and Payment Habits: The Role of 'Typical' and 'Specific' Time Frames in Survey Questions." In *Improving the Measurement of Consumer Expenditures*, edited by Christopher Carroll, Thomas Crossley, and John Sabelhaus, chapter 15. NBER.

Battistin, Erich and Mario Padula. 2015. "Survey Instruments and the Reports of Consumption Expenditures: Evidence from the Consumer Expenditure Surveys." *Journal of the Royal Statistical Society, Series A* 179(2):559–581.

BHPS. Various Years. "British Household Panel Survey." https://www.iser.essex.ac.uk/bhps.

Blair, Edward and Scott Burton. 1986. "Processes Used in the Formulation of Behavioral Frequency Reports in Surveys." *American Statistical Association Proceedings of the Section on Survey Methods* pp. 481–487.

BMCS. Various Years. "Biennial Media Consumption Survey." http://www.cpanda.org/data/profiles/bmcs.html.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement ERror in Survey Data." In *Handbook of Econometrics*, edited by James J. Heckman and Edward Leamer, volume 5, pp. 3705–3843. Elsevier.

Bradburn, Norman M., Lance J. Rips, and Steven K. Shevell. 1987. "Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys." *Science* 236:157–161.

BRFSS. Various Years. "Behavioral Risk Factor Surveillance System." http://www.cdc.gov/brfss/.

Brzozowski, Matthew, Thomas F. Crossley, and Joachim K. Winter. 2017. "A Comparison of Recall and Diary Food Expenditure Data." *Food Policy* 72:53–61.

CES. Various Years. "Consumer Expenditure Survey." http://www.bls.gov/cex/.

Chatzitheochari, Stella, Kimberly Fisher, Emily Gilbert, Lisa Calderwood, Tom Huskinson, Andrew Cleary, and Jonathan Gershuny. 2018. "Using New Technologies for Time Diary Data Collection: Instrument Design and Data Quality Findings from a Mixed-Mode Pilot Survey." *Social Indicators Research* 137 (1):379–390.

Clarke, Philip M., Denzil G. Fiebig, and Ulf-G. Gerdtham. 2008. "Optimal Recall Length in Survey Design." *Journal of Health Economics* 27:1275–1284.

Crossley, Thomas and Joachim Winter. 2014. "Asking Households about Expenditures: What Have We Learned?" In *Improving the Measurement of Consumer Expenditures*, edited by Christopher Carroll, Thomas Crossley, and John Sabelhaus, chapter 2. NBER.

DCPC. Various Years. "Diary of Consumer Payment Choices."

Deaton, Angus and Margaret Grosh. 2000. "Consumption." In *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten years of LSMS Experience*, edited by Margaret Grosh and Paul Glewwe, chapter 17. The World Bank.

Eisenhower, Donna, Nancy A. Mathiowetz, and David Morganstein. 1991. "Recall Error: Sources and Bias Reduction Techniques." In *Measurement Errors in Surveys*, edited by Paul P. Biermer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. Wiley.

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27:2865–2873.

Gelman, Andrew and Donald Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science* 7:457–511.

Greaves, Stephen, Adrian Ellison, Richard Ellison, Dean Rance, Chris Standen, Chris Rissel, and Melanie Crane. 2015. "A Web-Based Diary and Companion Smartphone App for Travel/Activity Surveys." *Transportation Research Procedia* 11:297–310.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York, New York: Wiley.

Groves, Robert M., Don A. Dillman, John L. Eltinge, and Roderick J.A. Little. 2001. *Survey Nonresponse*. New York: Wiley.

Hurd, Michael and Susann Rohwedder. 2009. "Methodological Innovations in Collecting Spending Data: The HRS Consumption and Activities Mail Survey." *Fiscal Studies* 30:435–459.

Jonker, Nicole and Anneke Kosse. 2009. "The Impact of Survey Design on Research Outcomes: A Case Study of Seven Pilots Measuring Cash Usage in the Netherlands." DNB Working Paper 221. De Nederlandsche Bank .

Kemsley, William F. F. and J. L. Nicholson. 1960. "Some Experiments in Methods of Conducting Consumer Expenditure Surveys." *Journal of the Royal Statistical Society, Series A* 123(3):307–328.

Marini, Margaret Mooney and Betth Anne Shelton. 1993. "Measing Household Work: Recent Expereince in the United States." *Social Science Research* 22:361–382.

McKenzie, John. 1983. "The Accuracy of Telephone Call Data Collected by Diary Methods." *Journal of Marketing Research* 20:417–427.

Means, Barbara, Gary E. Swan, Jared B. Jobe, James L. Esposito, and Elizabeth F. Loftus. 1989. "Recall Strategies for Estimation of Smoking Levels in Health Surveys." Paper for the American Statistical Association Meetings.

Menon, Geeta. 1994. "Judgments of Behavioral Frequencies: Memory Search and Retrieval Strategies." In *Autobiographical Memory and the Validity of Retrospective Reports*, edited by Norbert Schwarz and Seymour Sudman, pp. 161–172. Springer-Verlag.

Neter, John and Joseph Waksberg. 1964. "A Study of Response Errors in Expenditure Data from Household Interviews." *Journal of the American Statistical Association* 59:18–55.

NHIS. Various Years. "National Health Interview Survey." http://www.cdc.gov/nchs/nhis.htm.

NSSO Expert Group on Sampling Errors. 2003. "Suitability of Different Reference Periods for Measuring Household Consumption: Result of a Pilot Study." *Economic and Political Weekly* 37(4):307–321.

Nusser, Sarah M., Nicholas K. Beyler, Gregory J. Welk, Alicia L. Carriquiry, Wayne A. Fuller, and Benjamin M. N. King. 2012. "Modeling Errors in Physical Activity Recall Data." *Journal of Physical Activity and Health* 9:56–67.

PSID. Various Years. "Panel Study of Income Dynamics." http://psidonline.isr.umich.edu/.

Rockwood, Todd. 2015. "Assessing Physical Health." In *Handbook of Health Survey Methods*, edited by Timothy P. Johnson, chapter 5. Hoboken, New Jersey: John Wiley and Sons.

SCA. Various Years. "Survey of Consumers." http://www.sca.isr.umich.edu/.

Schmidt, Tobias. 2011. "Fatigue in Payment Diaries - Empirical Evidence from Germany." Discussion Paper Series 1: Economic Studies No 11/2011. Deutsche Bundesbank .

Shephard, RJ. 2003. "Limits to the Measurement of Habitual Physical Activity by Questionnaires." *British Journal of Sports Medicine* 37 (3).

Siemieniako, Dariusz. 2017. "The Consumer Diaries Research Method." In *Formative Research in Social Marketing: Innovative Methods to Gain Consumer Insights*, edited by Krzysztof Kubacki and Sharyn Rundle-Thiele, pp. 53–66. Singapore, Malaysia: Springer Singapore.

Silberstein, Adriana R. and Stuart Scott. 1991. "Expenditure Diary Surveys and Their Associated Errors." In *Measurement Errors in Surveys*, edited by Paul P. Biermer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. Wiley and Sons.

Strube, Gerhard. 1987. "Answering Survey Questions: The Role of Memory." In *Social Information Processing and Survey Methodology*, edited by Hans-J. Hippler, Norbert Schwarz, and Seymour Sudman, pp. 86–101. Springer-Verlag.

Sudman, Seymour and Norman M. Bradburn. 1973. "Effects of Time and Memory Factors on Response in Surveys." *Journal of the American Statistical Association* 68(344):805–815.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Thomas, Neal, Ofer Harel, and Roderick J.A. Little. 2016. "Analyzing Clinical Trial Outcomes Based on Incomplete Daily Diary Reports." *Statistical Medicine* 35 (17):2894–2906.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133 (5):859–883.