Survey Research Methods (2017)
Vol. 11, No. 3, pp. 219-233
doi:10.18148/srm/2017.v11i3.6794

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Assessing Potential Errors in Level-of-Effort Paradata using GPS Data

James Wagner
Survey Research Center
University of Michigan
Ann Arbor, MI, U.S.A.

Kristen Olson
Department of Sociology
University of Nebraska-Lincoln
Lincoln, NE, U.S.A.

Minako Edgar
Survey Research Center
University of Michigan
Ann Arbor, MI, U.S.A.

Surveys are a critical resource for social, economic, and health research. The ability to efficiently collect these data and develop accurate post-survey adjustments depends upon reliable data about effort required to recruit sampled units. Level-of-effort paradata are data generated by interviewers during the process of collecting data in surveys. These data are often used as predictors in nonresponse adjustment models or to guide data collection efforts. However, recent research has found that these data may include measurement errors, which would lead to inaccurate decisions in the field or reduced effectiveness for adjustment purposes (Biemer, Chen, & Wang, 2013; West, 2013). In order to assess whether errors occur in level-of-effort paradata for in-person surveys, we introduce a new source of data – Global Positioning System (GPS) data generated by smartphones carried by interviewers. We examine the quality of the GPS data. We also link the GPS data with the interviewer-reported call records in order to identify potential errors in the call records. Specifically, we examine the question of whether there may be missing call records. Given the lack of a gold standard, we perform a sensitivity analysis under various assumptions to see how these would change our conclusions.

Keywords: Paradata; Measurement Error; In person surveys; call records

1 Introduction

Surveys are a critical resource for social, economic, and health-related studies of human populations. However, surveys are facing a crisis of rising costs and falling response rates (Brick & Williams, 2013; Presser & McCulloch, 2011). These difficulties threaten both the validity and the viability of survey data. Survey methodologists have been searching for methods to sustain and improve the quality of survey data. Central to these efforts are the unique survey form of 'big data' known as paradata. Paradata are data generated during the process of implementing a survey (Couper, 1998; Couper & Lyberg, 2005). Although the measurement process was the initial focus for paradata (Couper, 1998; Olson & Parkhurst, 2013; Yan & Olson, 2013), increasingly, interest has been in paradata related to the recruitment and participation process (Couper & Wagner, 2011; Eckman, Sinibaldi, & Mantmann-Hertz, 2013; Kreuter, 2013; Wagner, 2013). These data about the recruitment and participation process have been characterized as "level-of-effort paradata" (Biemer et al., 2013). Level-of-effort paradata are generated by interviewers in the form of "call records," records of each attempt made to contact and interview households. Each record typically includes the time, date, mode of call (e. g., telephone or in-person), and an outcome or result code. For in-person surveys, these data provide limited information about what interviewers are actually doing in the field, including travel patterns.

Level-of effort paradata are important for evaluating and improving the quality of survey data (for a summary see Kreuter, 2013). Many studies use paradata to improve survey design, such as in the implementation of responsive designs (Groves & Heeringa, 2006) or to identify cases that may require a different recruitment protocol (Peytchev, Baxter, & Carley-Baxter, 2009). Further, these data have also been made available to improve weighting and imputation procedures for users of several major publicly-available datasets, including the National Health Interview Survey, the American National Election Studies, and the European Social Survey. Other surveys have included summary measures of these level-of-effort paradata on public use files, including the to-

Contact information: James Wagner, Michigan Program in Survey Methdology, P4053 ISR, 426 Thompson St., Ann Arbor, MI, USA (email:jameswag@umich.edu)

tal number of call attempts to a sampled case and whether the respondent had ever refused.

Despite the ubiquity of level-of-effort paradata in surveys, the quality of these paradata has received little attention. Errors in paradata may affect how one allocates effort in the field (e. g. Calinescu, Bhulai, & Schouten, 2013), leading to suboptimal deployment of resources. Measurement errors in paradata may reduce their effectiveness for adjustment purposes (Biemer et al., 2013; West, 2013). However, previous research does not explore the extent to which these errors may occur.

One difficulty in evaluating the quality of level-of-effort paradata is identifying a measure of quality. In this paper, we review the scant existing literature on the quality of level-of-effort paradata for in-person surveys. We then employ a relatively new technique – collection of GPS data in real time during the field period via a smartphone GPS application. These GPS data are not without errors. Therefore, we examine the quality of these data. We also have surveyed interviewers on both travel behavior and potential errors in the generation of paradata. Finally, we compare the level-of-effort paradata to the GPS data. Although the GPS data have errors, they may be useful in identifying patterns of behavior that are meaningfully related to survey outcomes that are not documented in the paradata. In particular, we examine the following questions:

1. What is the quality of GPS data collected during the process of data collection?

2. What insights can be gleaned from the comparison of GPS data with call record paradata?

This paper provides the first insight into the challenges and opportunities that arise when using real-time GPS data collection. We use these GPS data to evaluate the magnitude of measurement errors in paradata.

2 Background

For in-person surveys, interviewers are trained to generate call records for each contact attempt on laptops, tablets, or smartphones. These computerized sample management systems may automatically record the date, time, and (in certain situations) the result code (e.g., completed interview) of each call attempt. Interviewers may edit these fields, and may also make decisions about which kind of actions warrant generating a call record. These data are called level-of-effort paradata.

Level-of-effort paradata are used by virtually every survey organization to manage data collection efforts. For example, paradata are used to estimate response, contact and refusal rates; to identify the optimal day and time of visits to sampled units (Durrant, D'Arrigo, & Steele, 2011; Kreuter, 2013; Kulka & Weeks, 1988; Weeks, Jones, Folsom Jr., & Benrud, 1980, 1987); and to evaluate whether refusal conversion efforts are needed (Beullens, Billiet, & Loosveldt, 2010; Dutwin et al., 2015). Responsive or adaptive survey designs (Groves & Heeringa, 2006) draw extensively on level-of-effort paradata to inform design decisions, manage surveys (Kirgis & Lepkowski, 2013) and intervene during data collection (Kreuter, Mercer, & Hicks, 2014; Peytchev et al., 2009; Wagner et al., 2012). Level-of-effort paradata from existing studies are also used to design future data collection efforts (e. g. Calinescu et al., 2013; Luiten & Schouten, 2013; Peytchev, Rosen, Murphy, & Lindblad, 2010). In practice, almost every survey relies on these data to monitor data collection and make design decisions in the field.

Level-of-effort paradata are also used for nonresponse adjustment. There is a relatively large literature about adjustment strategies based upon these data (Alho, 1990; Beaumont, 2005; Biemer et al., 2013; Biemer & Link, 2007; Drew & Fuller, 1980; Groves & Couper, 1998; Kreuter & Kohler, 2009; Potthoff, Manton, & Woodbury, 1993; Wood, White, & Hotopf, 2006) in which the values for survey outcome variables for nonrespondents are assumed to be similar to those respondents who required extensive recruitment efforts. Research using level-of-effort paradata for adjustment purposes assumes that they are error-free. Yet few direct examinations of potential errors that occur in level-of-effort paradata exist.

There are some reports of interviewer errors in paradata. Biemer et al. (2013) surveyed field interviewers who reported that they would not generate a call record if they drove by a housing unit and decided that no one was home, leading to an underreporting bias in the call records. This practice leads to an upward bias in estimates of contact rates because known noncontacts are systematically omitted from the call records. Interviewers in the same survey reported that they would keep cases near the maximum number of calls "alive" by not recording call attempts. This practice would have a similar biasing impact on estimates of contact and cooperation rates. Without knowledge of specific, empirical error rates, Biemer and colleagues used simulation methods to explore the potential impact of these errors in the number of call attempts on nonresponse adjustments. They conclude that even small errors can generate relatively high rates of bias in adjusted estimates; a mere (simulated) 5% underreporting in the number of calls resulted in a bias in the adjusted estimates of about 19% (Biemer et al., 2013, p. 165).

Interviewer mistakes are another potential source of errors. Early investigations of CAPI studies found that data entry errors occur during computer-assisted interviews, and are affected by the method of data entry (Baker, Bradburn, & Johnson, 1995; Couper & Groves, 1992; Dielman & Couper, 1995). Additionally, despite instructions and training, interviewers often fail to follow protocols for interviewing, including failing to read questions exactly as written between 3% and 73% of the time (Ongena & Dijkstra, 2006). On the other hand, these administration behaviors can be improved through training (Billiet & Loosveldt, 1988; Fowler, 1991). Further, studies of *respondents* have found that they make errors when asked to recall events that occurred in the past (Tourangeau, Rips, & Rasinski, 2000), but this concept has not been applied to the creation of interviewer-generated paradata. These same mechanisms may affect paradata generation. For example, paradata may be incorrect due to data entry errors or problems with recall when records are generated after a lengthy interval (e. g. several hours later).

There are a few evaluations of the quality of paradata using measures created from the paradata themselves, that is, an *internal* evaluation of paradata quality. Bates, Dahlhamer, Phipps, Safir, and Tan (2010) examined internal consistency in paradata across three large U. S. federal surveys. They found that call attempts that had a noncontact code were more likely to have call records generated at a time other than when the attempt actually occurred, suggesting that reporting errors may occur differentially across cases and outcomes.

The existing evidence suggests that field interviewers may underreport call attempts for a variety of reasons – data entry errors, to save time, through recall errors, or to prevent active cases from being closed – and that these errors may occur at different rates across different types of call outcomes (e. g., contacts vs. noncontacts). Quantifying this underreporting is a difficult task. One possible strategy would be to "shadow" interviewers and generate a second set of records to be compared to those of the interviewer (e. g., see Kalsbeek, Botman, Massey, & Liu, 1994). Such studies are expensive and, hence, rare.

One source of data about the survey recruitment and participation process that has yet to be considered comes from equipping interviewers with global positioning system (GPS) devices (Nusser, 2007). GPS devices can measure location (latitude and longitude), elevation, speed, direction, date, time, the source of the measurement (satellite or cell tower) and an indicator of accuracy of the measurement (e. g., dilution of precision measures) at regular time intervals. This creates a dataset of "GPS points" for each point in time. These GPS data have been used in the development of address frames (Cecchi & Marquette, 2012; Dekker, English, Winfrey, & Seeger, 2013; Levinsohn et al., 2010; Morton et al., 2007; Seeger, 2011), to capture the location of an interview as a method of interview validation and falsification detection (e. g. Cecchi & Marquette, 2012; Ellis, Sikes, Sage, Eyerman, & Burke, 2011; Haddaway, 2013; Keating, Loftis, McMichael, & Ridenhour, 2014; Sikes, 2009; U.S. Census Bureau, 2014), to help interviewers plan travel routes in field studies (Nusser & Fox, 2002), and for basic monitoring of field data collection (Kurkowski, 2013).

The GPS data are not without errors themselves (Olson & Wagner, 2015). First, GPS data can be missing (Lemmens, 2011) due to technical problems or because of interviewer failure. Technical problems occur when the GPS device fails to pick up a signal, picks up a poor signal, or when the telephone battery dies. In a pilot test for a travel survey, respondents used a high quality GPS-only wearable device in a single metropolitan area, Toronto (Chung & Shalaby, 2005), resulting in a missing data rate of 21.5%, due primarily to device failure. Further, different quality measurements may be produced depending upon whether satellites or cell towers are used for GPS measurement. In a study of GPS measurements using an iPhone, Zandbergen (2009) found that the device failed to identify up to 12.3% of known locations, depending on the type of signal used for the GPS measurement. The interviewer might forget to turn on the GPS logging application. For example, in a recent review of physical activity studies where *respondents* are asked to carry GPS devices (Krenn, Titze, Oja, Jones, & Ogilvie, 2011), missing data rates ranged from 2.5% to 92%, with one of the strongest predictors being the length of the study period (a long study took four or more days).

Second, even when a signal is not completely blocked, it is still possible for GPS data to have measurement error (Chung & Shalaby, 2005). The major reasons for measurement error in GPS data are inadequate satellite signals, errors in transmitting from the satellites to the ground, or the need to use the less accurate cell phone towers (Goodchild et al., 2007; Lemmens, 2011; Zandbergen, 2009). Different devices may result in different levels of error – cell phones have higher rates of signal loss and less accurate measurement than specialized GPS devices (Wu et al., 2010).

Linking GPS data with call record data is a difficult task. Others have tried to link data from travel surveys to GPS data from respondents. Mavoa, Oliver, Witten, and Badland (2011) found that using a computer algorithm that compares sequences from travel diaries to sequences in the GPS data led to linkages that were of sufficient quality for analysis purposes compared to manual linkages, but at lower costs. Deriving characteristics of the trip, such as the purpose and mode of travel (e. g., car, bike, train, foot) is another challenge when analyzing GPS data. Bohte and Maat (2009), for example, use a set of rules to derive the purpose and mode of trips from GPS data. When compared to a daily web survey of the same persons from whom they collected the GPS data, the rule-based approach provided reliable estimates of purpose and mode. These studies provide evidence that GPS data can be reliably used for these types of purposes.

In sum, neither the call records nor the GPS measurements are error-free. Nevertheless, linking the data about specific interviewer trips to sampled neighborhoods and housing units available from the two sources will illuminate strengths and weaknesses in each data source. Such a comparison may shed light on survey interviewer travel patterns, identify potentially missing data in the level-of-effort paradata, and illuminate additional detail not available in the paradata. This is a first step toward understanding and quantifying errors in level-of-effort paradata, with longer term implications on survey management and adjustment methods.

3 Data

The data come from the National Survey of Family Growth Continuous 2011-2019 (NSFG). The NSFG interviews females and males between the ages of 15 and 44 about fertility-related topics. The NSFG is an in-person survey, with a continuous area probability sample design that rotates Primary Sampling Units (PSUs) annually. A new independent sample of housing units is released four times a year (quarterly); each sample is in the field for 12 weeks. For any given year, the NSFG has 40 to 45 female interviewers on staff. The interviewers work in 35 primary sampling units with over 450 unique area segments, with approximately 20,000 housing units sampled each year. Each area segment contains a minimum of 75 housing units and the average number of housing units in an area segment is 179. The sampling frame is developed from commercially available address lists. These lists are checked in the field by interviewers. A sample is then selected from this frame of housing units. Cumulatively, these sampled units receive more than 100,000 call attempts annually. We will examine data from quarters 1 through 8. During these quarters, the response rate ranged from 71% to 76% (AAPOR RR4), with a total of 39,494 sampled housing units.

Interviewers have convertible tablet computers with the University of Michigan Survey Research Center sample management system, SurveyTrak. Interviewers keep call records for attempts to contact housing units, screen for eligible respondents, randomly select respondents from the household, and conduct an interview. The sample management system data are entered into SurveyTrak in the field, with the current time and date automatically loaded into each call record. Interviewers are able to edit these times and dates for the most recent call attempt. The remaining result code, mode, and call notes are recorded by the interviewer.

Interviewers for the NSFG also were equipped with web-enabled smartphones. These Android-based smartphones (Motorola Atrix 2) included an application or “app” (GPS Logger for Android devices) that recorded their location (latitude and longitude), time, date, speed, direction of travel, altitude, source of measurement (satellite vs. cell towers), and two indicators of the accuracy of these data (the number of satellites used in the measurement and the Horizontal Dilution of Precision, a measurement error indicator). The app was configured to take measurements every 60 seconds. The interviewers were instructed to activate the application when they began their shift and stop the application when their shift was complete. Olson and Wagner (2015) give a more detailed description of how these data are collected, interviewer compliance with the request to use the app while working, and whether the use of the app altered the behavior of interviewers.

We have a total of 1,943,764 individual GPS points in the 7,168 interviewer-days of recordings. The average file has 271.2 GPS points, covering an average distance of 149.5 miles (median distance = 51.98 miles) over an average of 6.8 hours (median time = 5.8 hours).

The other data source comes from web surveys of and debriefings with the interviewers. Three web surveys were conducted with the NSFG interviewers during summer 2012 (n=29, AAPOR RR1=62%), summer 2013 (n=25, AAPOR RR1=71%), and summer 2014 (n=23, AAPOR RR1=60%). We also conducted hour-long debriefing sessions with ten interviewers between March 1, 2012 and July 23, 2012. These debriefings were based on a purposive sample of interviewer-days that were selected because of apparent discrepancies between the call records and the GPS data or longer than average travel distances.

4 Sampling and Linking Call Records and GPS Data

Call record and GPS data were collected in two separate systems, and thus need to be linked together, a nontrivial task. There are several ways in which this linkage can be done. Our purpose is to evaluate the completeness of call records. However, there may be missing data from the GPS data as well. Missing GPS data can occur because the interviewer forgets to start the app or because the phone does not have access to either satellites or cell towers. Further, the linkages can be incorrectly made.

Since the linkage was computationally intensive, we drew a sample of interviewer-days. We selected a simple random sample of 179 interviewer-days from all interviewer-days in Q2 using the call records to identify all days when an interviewer worked. Of these 179 days, 101 (56.4%) had GPS files available. These 101 interviewer-days had a total of 917 call attempts to a total of 655 housing units as recorded in the call records. The GPS files for these 101 days had a total of 2,825 routes that passed active sampled housing units that needed to be processed and merged to the call records.

The sample we selected allows us to infer to the population of interviewer-days during Q2 of the NSFG. Our analysis is further restricted in that we have GPS data for a subset of interviewer-days in our sample. This subset was chosen by the interviewers via their compliance with the request to turn the app on and off. Olson and Wagner (2015) examine how days with differ from days without GPS files. They found that GPS logging was more likely to occur on interviewer-days with more call records.

We use a deterministic linkage procedure, described in Table 1. There are two dimensions which need to be incorporated in the linking process: time and space. Unfortunately, the measurement on both of these dimensions can have errors. Even when accurately measured, recorded GPS points and geocoded location of sample housing units are rarely at

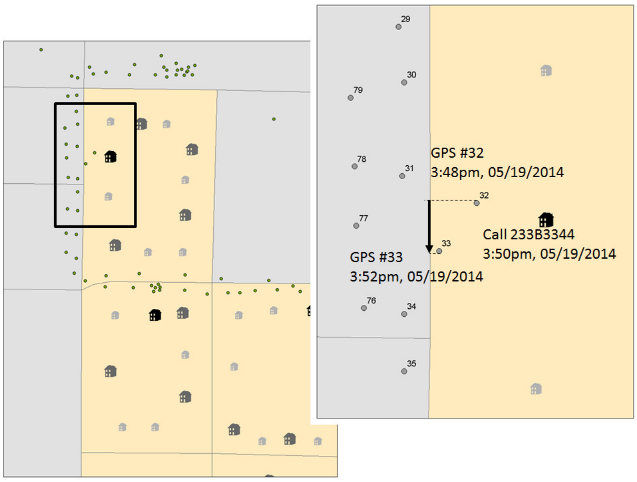

Figure 1. Graphical Representation of Merging Call Records with GPS Data

identical latitude and longitude measurements. Therefore, a maximum distance needs to be identified for a GPS point to be considered in the same location as an address in the call records. Additionally, using time as the linkage criterion is challenging because the time on the clock can be different on the tablets and smartphones. This issue is exacerbated when GPS points are missing due to poor reception of satellite signals. Finally, the call records may be generated several minutes after the call attempt and in a location other than the sampled housing unit. Thus, the deterministic linkage procedure uses a combination of distance and time for identifying links.

Given the complexity of the procedure, we describe how this linkage was done in detail, using an example map to illustrate the process. We also provide additional details in an the appendix. We start with the date and identify call attempts and GPS points that occur on the same date recorded in both systems. Only days for which we had *both* call records and GPS files were linked.

Figure 1 gives a graphical representation of how these data are linked using simulated data. The figure on the left shows the GPS points and sampled housing units. The figure on the right zooms in on a few points.

Focusing on left side of Figure 1, the larger-size house icons represent geocoded sampled housing units and the smaller-size house icons are unsampled housing units. The larger icon black houses are those sampled houses that have a call record with a date, time, and outcome code for that day. The larger icon gray houses are sampled units that do not have a call record or are not active. These dates and times were either generated automatically by the interviewer's laptop or were edited manually by the interviewer. The gray dots represent the GPS points recorded by the interviewer's smartphone.

We started by sorting the GPS points (the gray dots) using the date and time stamps on each point. Measurement error in the recorded GPS points leads to variation in placement of points, as illustrated in Figure 1. Each GPS point in our sample was "snapped" to the nearest road (street, highway, etc.). Then routes between GPS points are created. This assignment is represented in the figure by the dashed lines – for example, points 32 and 33 were snapped to the nearest road along these dashed lines. Once snapped to the nearest road, a route along these GIS-identified roads between the two points was identified following the temporal order of the GPS points. The route follows the nearest roads to travel between two points. In Figure 1, the arrow represents the inferred route taken between points 32 and 33. We assigned the time stamp of the last point on the route as the time of the inferred route.

In the next step, call records were linked to these routes. As noted earlier, we do not expect a call record should be available for every route. Call records were similarly sorted by date and time of the call attempt and sampled housing units were geocoded. We then selected the route with the time stamp immediately preceding the call record time. If all the routes were later in time than the call record, we merged the call record with the route with the earliest time stamp. Out of the n=101 interviewer-days with GPS files, we created 2,825 records of inferred routes that passed in front of sampled housing units that had not been finalized as of the date of the route. For those 101 interviewer-days, there were a total of 917 unique call records. If there was a call record associated with that sampled unit for that day, then it was attached to the route. It was possible for multiple routes to pass the same housing unit. For example, if an interviewer walked up and then down the same street, this would generate two routes past any sampled housing unit on that street. This meant that a single housing unit could be associated with multiple inferred routes. Each call record was linked to only one route using the method described.

Figure 1 provides an example of the linking process. The right figure in Figure 1 is a zoomed in view of the part of the left figure outlined by the black box. In Figure 1, there is a call record associated with the black housing unit. There are two routes that pass sampled housing unit 233B3344: the route associated with travel between points 32 and 33 and the route between points 77 and 78. Since the route between points 32 and 33 immediately follows the time of the call record, we associate this route with the call record.

Finally, we determined whether each sampled unit was active on the day of the route. We did this by comparing the date of the final result code from each housing unit to the date of the route. If it was earlier than the date of the route, we treated that sampled unit as having been finalized before the day of the route and, therefore, not available for calling. Of the 2,825 routes associated with active sampled housing

Table 1
Description of GPS and Call Record Linkage Procedure

GPS Processing

Step 1: Sort GPS points using the GPS date and time stamps within each interviewer-day.
Step 2: "Snap" GPS points to nearest road.
Step 3: Identify route along road between two1 points adjacent in time.
Step 4: Assign last time stamp for the route as the "time of trip."

Call Record Processing

Step 1: Identify subset of interviewer-days for which call records and GPS points recorded
Step 2: Geocode location of all sampled housing units.
Step 3: Identify "active" sampled housing units on that interviewer-day.

GPS and Call Record Linkage

Step 1: Calculate difference in distance (meters) between all inferred routes on a given interviewer-day and geocoded location of active sampled housing units.
Step 2: If the calculated difference in distance >1000 m, exclude inferred route as possible "link" to call records.
Step 3: If the calculated difference in distance < 1000 m, compare the street name for the sampled active housing unit to the street name of three closest inferred route.
Step 4: If the street name for the route is the same as the street name for the sampled active housing unit, identify as "potential link."
Step 5: If the street name for the route is not the same as the street name for the sampled active housing unit, then if there are multiple routes having same street name evaluate whether the closest point for a house to the nearest road is on an inferred route. If it is, identify as a "potential link."
Step 6: Among all of the "potential link" inferred routes, select the route with the GPS time stamp immediately before the call record time. This is a "linked route."
Step 7: If all of the GPS time stamps are after the call record time, select the route with the earliest time stamp. This is a "linked route."
Step 8: Identify the number of call records that do not have a linked inferred route.

units, 2,235 (79.1%) did not have a call record. The final dataset included three types of records (Table 2): 1) routes that passed a nonfinal sampled housing unit for which there was a linked call record (n=590), 2) routes for which there was no call record (n=2,235), and 3) call records for which there was no route (n=327).

We assess the sensitivity of our conclusions to potential errors in the linkage by examining different subsets of linked records defined by time, speed, and distance. Overall, 20.9% of routes have a linked call record. We do not expect a call record for every route. However, every route that passes an active, sampled housing unit gives the interviewer an opportunity to observe potentially useful information. We might also assess the probability that there is a linkage for the call records. Overall, 35.7% of call records had a linked route. We would expect a route to be linked to every call record. These errors point to potential issues with the GPS data. Initial analyses indicate that routes are not systematically missing for any single type of call outcome.

We assess the sensitivity of our analyses to a number of factors. First, the interviewer may have been on their way to an appointment, and thus would not be able to stop at a sampled housing unit if they saw someone at home. We expect that there will be a higher probability of a linked route when the call attempt is not being made to an appointment.

Apartment complexes present a special problem for our method. Often, these units are geographically close to each other, including being located above or below other sampled units. Some apartment complexes share a common street, such that travel between apartments might be missed by our process of creating routes. Therefore, excluding apartments from the analysis might improve the overall accuracy of the linkages in the analysis set.

The next subset analysis is based upon the speed of travel. The interviewer may have been driving past active, sampled housing units and did not notice whether there seemed to be anyone at home. Therefore, the speed of travel along the route may also be an indication of whether there is any data missing. We expect that there will be a higher probability of a linked route when examining only call records with slower speeds.

Finally, some of the routes that we generate may not reflect actual interviewer travel. Due to errors in the GPS measurement, a stationary person can generate two different measurements of their location. These errors are usually fewer than 10 meters. Therefore, focusing on longer routes can

Table 2
Linkage Results for GPS and Call Record Data

	n	%	% of call records	% of routes
Total number of call records	917		100.0	
Total number of routes	2,825			100.0
Call record and GPS route	590	18.7	64.3	20.9
Call record, no GPS route	327	10.4	35.7	
GPS Route, no call record	2,235	70.9		79.1
Total	3,152	100.0		

reduce the possibility that the "route" is actually due to this sort of measurement error.

5 Results

In this section, we will first discuss the quality of the GPS data and our ability to link them to the call record data. Second, we will discuss interviewer reports of errors they make in the process of generating paradata. Then, from the combined call records and GPS data, we will look at measures of potential errors in the call records. Finally, we will use several measures to assess the sensitivity of our results to the quality of the linkage.

5.1 GPS Data

The GPS data are useful when they are (1) high quality and (2) can be successfully merged with the call record information. We start by examining the quality of the GPS data with respect to potential technological implementation deficiencies.

An initial, if not very discriminating, measure of quality is whether the GPS measurements were estimated using satellite signals or through cell phone network towers (Zandbergen, 2009). Of the 1.9 million GPS points, 69.4% were recorded via the higher-quality satellite measurements (with an average of 7.61 satellites used for the measurement), and 30.6% were recorded via the lower-quality cell network.

A second internal measure of quality is the Horizontal Dilution of Precision (HDOP). This measure provides a more detailed assessment of the precision of the measurement. This measure was not implemented in the app until after the survey had been in the field for a year. We have HDOP data for 230,234 GPS points (Table 3). By this measure, only 21.6% of the points meet the accepted criterion for "good quality" measurement – having HDOP values less than or equal to 5 (Chung & Shalaby, 2005; Rempel & Rodgers, 1997; Stopher, FitzGerald, & Zhang, 2008). About a fifth of the points have HDOP values above 20, considered to be only "poor" measurements (Piras & Cina, 2010). Given the use of smartphones for this measurement and interviewers being in cars or respondent's homes, it is likely that the GPS device did not have good positioning of the antenna (being in a computer bag, purse, or pocket) or have a clear view of the sky (being in a respondent's home), increasing HDOP values.

Table 3
Horizontal Dilution of Precision (HDOP) measurements from GPS Logger app, NSFG Q1-Q8

HDOP	Quality	N	%
0-1	Ideal	8,612	3.7
1.01-2.0	Excellent	9,557	4.2
2.01-5.0	Good	31,647	13.8
5.01-10.0	Moderate	56,894	24.7
10.01-20.0	Fair	71,030	30.9
20.01+	Poor	52,494	22.8
Total		230,234	100.0

Another data quality measure for the GPS measurements is the length of time between the points. The application was set to take this measurement every 60 seconds; a gap of more than a minute thus indicates a failure to find a satellite signal. Figure 2 shows the distribution of the lag times between points. The median length of time for all the points is 65 seconds. More than 19% of points have more than 120 seconds elapsed between them. The median elapsed time between points is equivalent for GPS measurements taken by satellite and those taken over the cell network (65 seconds), but the means differ substantially. Measurements taken by satellite are 125 seconds apart, on average, compared to 91 seconds apart for those recorded via cell towers.

Although the GPS data are not a gold standard, they provide sufficient detail for our purposes. That is, the available GPS data allow us to describe general patterns of interviewer travel and identify occasions when interviewers were near sampled housing units without making a call record. There is missing data in the GPS data, but observing those data could only change the magnitude of our findings regarding underreporting of call records, not whether this under-reporting occurs.

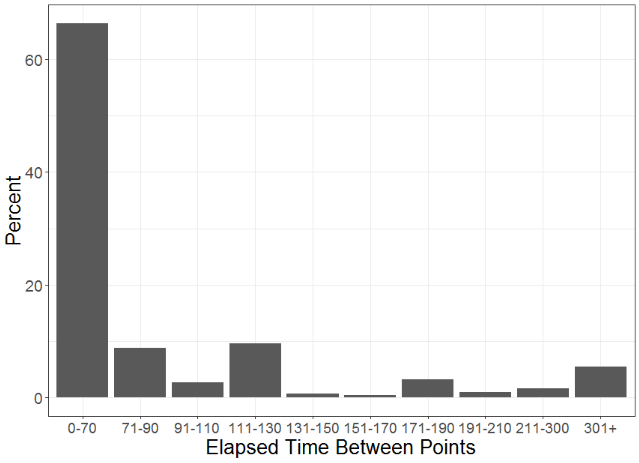

Figure 2. Elapsed Time in Seconds Between GPS Measurements Captured by GPS Logger App, NSFG Q1-Q8

5.2 Quality of Links between GPS and Call Record Data

We now examine the quality of the links between call records and GPS routes. There is no GPS route available for 35.7% of housing units with a call record on that interviewer-day. This might be an indication that the interviewer was never at the sampled housing unit for which the call was recorded. However, given the errors that we have seen with the GPS data and call records, it also may have been the result of missed GPS points (due to interviewer or technical failure), incorrect dates in the call records, or errors in the inferred route. Unfortunately, our data do not give us insight into which of these issues may have occurred.

The other 64.3% of housing units with a call record were merged to a route. We first assess the quality of these merges by comparing the time stamps on each data source. The median time difference between the GPS time stamps and the merged call record time stamps was 10.2 minutes. The 25th percentile was 3.2 minutes and the 90th percentile was 85.9 minutes. This indicates that the merge based first on geography and then on time generally worked, with minimal time discrepancies. An example of an outlying value occurs when the GPS time points run from 11am to 3pm, but the call record has a time of 7:30pm. This could indicate either that there are errors in the time recorded in the call records, or that the interviewer returned to a housing unit during the same day without turning on the GPS device.

Overall, although there are missing GPS data, for those interviewer-days where the GPS data are available, the linkages are quite good. In particular, 90% of these linked call records and GPS points have time stamps that are within 86 minutes of each other.

5.3 Interviewer Self-Reports of Errors in Paradata

Errors may occur in call records for a variety of reasons. In the debriefing interviews, interviewers explained that they often completed their call records in their car around the corner from a contacted sampled unit. Interviewers also told us that would wait until they had made several non-contact calls and then enter them into the sample management system, allowing them to focus on making attempts at households as efficiently as possible.

Additional evidence about errors was provided in the web surveys completed by the interviewers. Asked if they had made mistakes in generating paradata (on a five-point scale ranging from always to never), 47% of interviewers said they have ever recorded the wrong day, and 55% ever recorded the call record for the wrong address. Further, 88% reported that they would not make a call record if they walked passed a household and determined that no one was home without knocking on the door – a situation they are explicitly told in training should result in a call record.

Errors may also occur in call records when they are generated after the actual contact attempt was conducted, such as at the end of a shift or the next day. In these cases, it may be difficult for the interviewer to remember the times at which calls were actually attempted, and thus result in estimating the contact time, or that an attempt was made at all. Across the three web surveys, 83% of interviewers responded that they always or most of the time recorded call records immediately after the call attempt. On the other hand, 29% of interviewers reported completing call records at the end of the day at least once, and 4% reported completing their call records at the end of the week at least once.

5.4 Evidence of Under-Reporting in the Call Records

We now turn to an important issue: do the GPS data identify underreporting of call records? We will briefly discuss overall patterns of travel. Then we will examine this question using all of the linked data. After examining all of the data, we will then perform a sensitivity analysis where we focus on subsets of the data with indications of higher probabilities of correct linkages.

We begin with an examination of patterns of within-segment travel. There were 2,703 active sampled housing units in the sampled segments with GPS data. Table 4 shows the distribution of the number of times an active sampled housing unit was passed by an inferred route during a visit to an area segment. First, 61% of active sampled units in a sampled segment were not passed – that is, were not associated with an inferred route – on a particular interviewer-day. Not passing an active housing unit is due to a complex set of factors – the workload may be large, the shift may be short, or relatively more time was spent completing interviews at a few households.

Table 4
Distribution of Number of Times Active Sampled Housing Units are Passed Per Segment Trip

Number of Times HU is Passed	All Routes (%)
0	61
1	16
2	12
3	5
4+	6

Note: n=2703 active housing units

Second, there are non-linear patterns of travel in the GPS data, indicating back-and-forth movement. Among all active sampled units, 16% were passed one time, but 23% of all active sampled units were passed at least twice.

Using the inferred travel routes from the GPS points, we also can look at how many times a sampled housing unit – excluding those with finalized result codes – is passed without a call record being made. Of course, it is possible that the interviewer did not make an active decision about whether or not someone was home and no call record should be generated, but an opportunity for such a decision was possible.

We begin by looking at all the 590 routes merged to a call record and 2,235 routes for which there was no call record. As noted above, this means that there is a call record for only 21% of routes, and no call record generated for a striking 79% of all of the inferred interviewer routes that pass an active sampled housing unit (Table 5).

Interviewers are trained to make a call record if they pass a housing unit and determine that no one was home. The reason this rule is employed is to avoid creating a bias in estimated contact rates that would occur if these kinds of "noncontact" events were not recorded. The evidence that noncontacts are underreported is buttressed by the results from our survey of interviewers, where we found that 88% of interviewers would not generate a call record when they have determined that no one is home without knocking on the door.

The household may not have been observed by interviewers in the field, either due to appointments, being in an apartment building, or traveling quickly. We use information about these situations to evaluate the sensitivity of our results to various assumptions about the likelihood of a call record being necessary. This is accomplished by examining subsets of cases that are more likely to have correct linkages.

As a first step in this sensitivity analysis, we excluded routes where the time stamp on the GPS occurred 30 minutes or less before a call to a housing unit on that route with a scheduled appointment. This definition is imperfect; some cases have calls attempted during the 30 minutes prior to an appointment, and is thus somewhat conservative. With this exclusion the percentage of routes that went past a housing unit and did not have an associated call record goes down slightly to 76%, with the percentage of linked routes increasing slightly to 24%. Thus, appointment-keeping appears to be one of the reasons that a call record is not made, but the effect on estimated rates of observing a call record is modest.

It is possible that linkages are more difficult to make for apartment buildings than for single family homes. Apartments constitute 18% of the active sampled housing units included in the analysis. When we excluded all apartments, the results were essentially the same (80% not linked vs. 79% not linked overall). Therefore, linkages missed due to apartment buildings do not explain our findings.

We can also use the speed and distance of travel from the GPS data to further assess the sensitivity of our results. First, we exclude routes that were less than 15 m long as an indicator of potential measurement error in the GPS measurement. This has no effect on the linkage rate.

We note that an average human walking pace is about 1.3 meters per second (Pline, 1992). Bohte and Maat (2009) use an average speed of less than 2.8 meters per second and a maximum speed of less than 3.9 as an indication of foot travel. Interviewers were traveling more quickly along routes on which they did not make call records. The average speed of cases with a call record was 5.5 meters per second (about 12 miles/hour; median=1.6 m/s, or about 3.5 miles/hour). The average speed for cases with no call record was 12.0 meters per second (about 27 miles/hour; median=5.1 m/s, or about 11 miles/hour). The pattern is identical when excluding routes that appear to be on the way to an appointment. This speed suggests that interviewers were likely traveling too quickly on some of these routes to directly observe whether a sampled household member was at home. Figure 3 shows the proportion of routes that pass active sampled units for which there is no call record by the cumulative speed of travel. The speed is the maximum average speed of the included routes. For example, the point at 65% for 1.5m/s is the proportion for routes that had 1.5 m/s or lower average speed, a sharp decrease from the overall nonlinkage rate of 79%. From the figure, it is clear that although the proportion without call records increases as the average speed increases slightly, the range is limited, from about 65% to 71%. Therefore, interviewers driving by housing units and not noticing whether anyone is home because of the speed of travel is not the only explanation for our high estimates of nonlinkage.

As discussed above, interviewers can pass housing units several times. An alternative explanation for the lack of call records is that interviewers only generate one call record for each housing unit that is passed several times. While this may be true in some cases, among the 1,065 active sampled housing units that are ever passed by an inferred route (i. e. passed 1+ times), 31% do not have a call record recorded at all for the day. Thus, interviewers pass by housing units,

Table 5
Comparison of GPS data and call record data

	No call record			N (routes or
	%	SE	With call record (%)	housing units)
% of routes that pass a sampled housing unit				
All routes (row=100%)	79	2.4	21	2825
Excluding appointments (row=100%)	76	2.7	24	2486
Excluding apartments (row=100%)	80	2.5	20	2313
Excluding routes less than 15m (row=100%)	79	2.5	21	2692
% of active sampled housing units passed at least once by an inferred route				
All active sampled housing units (row=100%)	31	4.5	69	662
Excluding apartments (row=100%)	36	5.5	64	569

	No call record		With call record		
	Stat.	SE	Stat.	SE	N
Speed, all cases					
Average speed	12.0	2.9	4.7	0.9	2674
Median speed	5.1		1.7		
Speed, excluding appointment cases					
Average speed	11.4	2.5	4.7	0.9	2335
Median speed	5.2		1.7		

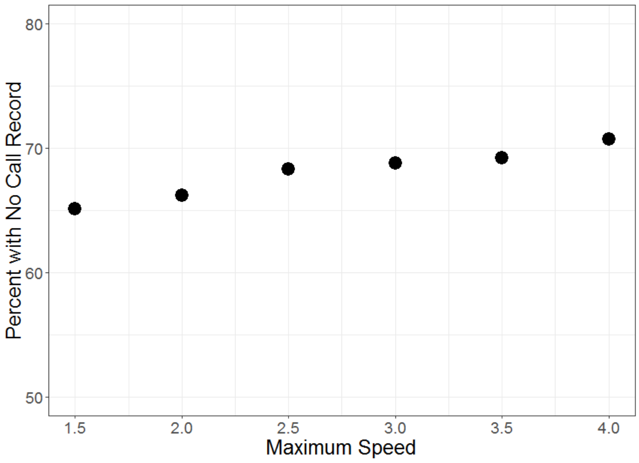

Figure 3. Percentage of Routes Passing an Active Sampled Housing Unit that do not have a Linked Call Record by Cumulative Maximum Speed of Routes

but do not record a call record roughly one-third of the time. Identifying the location of an apartment may be more difficult than a single family home. When we examine only non-apartments (e. g., single family homes), the percent of sampled housing units that are passed without a call record recorded for the day increases slightly to 36%.

Overall, the data suggest that call records are giving incomplete information about likely at-home rates. These results about the linkage rate show very limited sensitivity to any of the indicators of the quality of the linkage that we examined, with speed of travel yielding the largest effect on linkage rates.

6 Discussion

In this paper, we evaluated the quality of data collected from a GPS-enabled smartphone app and compared the GPS data with interviewer-recorded level-of-effort paradata. To our knowledge, this is the first study that has examined concordance of call record data with real-time GPS data. We identified that GPS data can provide unique insights into the movement of interviewers in the field, including that interviewers engage in back-and-forth travel. We also found that they move past some sampled housing units repeatedly. Other housing units (around 60%) are not passed at all on any given visit to a sampled segment. Many of these “passes” are undocumented in paradata, despite the fact that interviewers are trained to communicate this information through call records. Additionally, we found potential errors in both sources of data, and identified how the union of these two sources could provide new insights into interviewer travel behavior. This analysis suggests that further research is merited, since level-of-effort paradata are used for monitoring data collection and post-survey adjustment. Further, interviewers reported in debriefings and web surveys that they know that they make these sorts of errors, including potentially underreporting noncontact call attempts. The interviewer survey conducted by Biemer et al. (2013) reported similar results.

Our evaluation of GPS data indicated potential data quality issues that we hypothesize to be related to non-ideal circumstances for GPS measurement. Interviewers are likely to have their cell phone in a bag or in some other way obstructed and thus not able to obtain high quality GPS measurements. Lack of a clear view of satellites adds measurement error to the GPS points, as indicated by the low quality HDOP measurements, and irregular measurement time intervals, as indicated by long lapses between GPS points. We note that these problems are heightened because the interviewers were instructed to have the GPS app on at all times during work. Simply asking interviewers to “snap” GPS coordinates at the time of obtaining an interview would likely mitigate some of these precision problems at a particular housing unit, but detailed data about travel would not be possible in this approach. On the whole, these issues mean that GPS data cannot be considered a gold standard. However, they are accurate enough to allow researchers to gain insight into interviewer behavior in field surveys.

Finally, we compared call record data to GPS data. Two primary findings come from this analysis. First, the task of linking the GPS files and the call records is non-trivial, requiring substantial dedicated time by GIS professionals. Survey organizations that want to collect real-time travel data on interviewers for the purpose of monitoring or evaluating field work will need to build systems to assist with this linkage. This feasibility test did not permit such resources to be dedicated. Our second set of findings from this comparison has to do with interviewer behavior. We found that interviewers travel in nonlinear patterns through area segments, contrary to the models of within-area segment interviewer travel sometimes assumed by sample designs (e. g. Kalsbeek, Mendoza, & Budescu, 1983). Our survey of interviewers helped explain how these decisions are made and suggested that this nonlinear travel may be more efficient than a linear route that sought to minimize mileage. The GPS data also suggest that the call records contain incomplete information that may lead to overestimated contact rates. Interviewers travel past active sampled housing units a surprising number of times in their daily work travels, and this movement is not recorded in call records. To our knowledge, this has not previously been empirically described in an actual field study. This analysis did not reveal whether multiple trips to or by a housing unit on the same day is more or less effective than multiple contact attempts spread over different days of the week; future research taking advantage of both GPS and call record data will look at this question.

These findings are important for both survey management decisions and nonresponse adjustments. First, call records are used by virtually every survey organization to monitor field effort and guide decisions. As responsive or adaptive designs are becoming more prevalent (e. g. Groves & Heeringa, 2006; Miller, 2014), the accuracy of these data are increasingly important. Missing data on call attempts means that field effort is going unreported and interviewer decisions are not known to their supervisors or field managers, making these decisions less efficient and possibly less effective. Second, Biemer et al. (2013) simulation study demonstrated the important impact such call record errors could have on nonresponse adjustments. We suspect that these errors in call records may also influence estimates of coefficients and predictions in contact models (Wagner, 2013) and response propensity models used in responsive designs (Groves & Heeringa, 2006; Wagner & Hubbard, 2014; Wagner et al., 2012). The impact of the potentially missing data on both management decisions and on propensity models is not currently understood.

Reducing errors should also be a goal for survey organizations. There are at least three options. The first option is to train and monitor interviewers more carefully to improve level-of-effort paradata. Other studies have shown that training can improve interviewer performance with respect to administering the questionnaire (Billiet & Loosveldt, 1988; Fowler Jr. & Mangione, 1990) and reducing nonresponse (Groves et al., 1997; Groves & McGonagle, 2001). A problem with this approach is that it is very difficult (or impossible) to know when call records are missing, thus reducing the effectiveness of the training and monitoring. A second option is to use methods that passively collect data such as the GPS app used in this study. Given the difficulties we had merging the two disparate sources of data (call records and GPS data), more integrated systems may be required for this to be a practical solution. For example, can GPS-enabled devices be linked to sample management systems and passively record when an active sampled housing unit is passed? Can the interviewer be prompted by the sample management system when a sampled housing unit is passed? The generation of a call record could include the passive recording of the GPS location. The US Census Bureau recently conducted a test using a smartphone for sample management and interviewing for the US Decennial Census using an approach similar to this. When each call record was generated, the GPS location was recorded. In the test, if the phone’s latitude and longitude did not match that of a household where an interview was being conducted, the interviewer was asked if they were certain they were at the correct location (Walejko & Wagner, 2015). A third option is to ask interviewers to record some observations on each trip to an area segment. Did they drive through the neighborhood? Did they see evidence at any households that no one was home? These observations could be prompted either by call records generated in the area segment or by GPS devices that signal to the sample management system when the area segment is being entered or exited.

These options may have different costs. Asking interviewers to be more careful in the creation of call records may lead to additional costs. We found that 79% of the times that an interviewer passed a sampled housing unit, they did not generate a call record of this event. If we asked them to generate call records for all these events, we would be more than tripling the number of call records. Would these improved data be worth the costs? Bates et al. (2010) report that the median time to record a call record ranged between 43 and 55 seconds, depending on the survey. In the NSFG, an average quarter produces 22,000 call records. Producing $3 \cdot 22,000 = 66,000$ call records at 45 seconds each would require an additional 550 hours of interviewer time.

The costs of technological solutions are not free either. Integrating GPS technologies into existing sample management systems requires programming time. Further, these options need to be tested to see if they lead to changes in interviewer behavior (Olson & Wagner, 2015). The value of the data needs to be evaluated as well. Do improved call records reduce nonresponse bias of estimates? Does improved reporting of these data lead to more careful control of costs? Are field management decisions improved with better data? Understanding the cost and error implications of each of these decisions will be critical for the design of paradata. Further studies that aim to understand cost and error implications are needed to inform these decisions. The total survey error perspective should be a guiding principle for this research.

References

Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, *77(3)*, 617–624.

Baker, R. P., Bradburn, N. M., & Johnson, R. A. (1995). Computer-assisted personal interviewing: an experimental evaluation of data quality and cost. *Journal of Official Statistics*, *11(4)*, 413–431.

Bates, J., N., Dahlhamer, P., Phipps, A., Safir, & Tan, L. (2010). Assessing contact history paradata quality across several federal surveys. Proceedings of the American Statistical Association 2010 Joint Statistical Meeting.

Beaumont, J. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, *31(2)*, 227.

Beullens, K., Billiet, J., & Loosveldt, G. (2010). The effect of the elapsed time between the initial refusal and conversion contact on conversion success: evidence from the 2nd round of the European Social Survey. *Quality & Quantity*, *44(6)*, 1053–1065.

Biemer, P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176(1)*, 147–168.

Biemer, P. & Link, M. W. (2007). Evaluating and modeling early cooperator effects in RDD surveys. In J. M. Lepkowski, N. C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, . . . R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 587–617). Hoboken, New Jersey: Wiley.

Billiet, J. & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly*, *52(2)*, 190–211.

Bohte, W. & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17(3)*, 285–297.

Brick, J. M. & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645(1)*, 36–59.

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, *226(1)*, 115–121.

Cecchi, R. & Marquette, R. (2012). 2010 census: global positioning system evaluation. Edited by US Census Bureau. Washington, DC.

Chung, E. H. & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, *28(5)*, 381–401.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. Proceedings of the Survey Research Methods Section of the American Statistical Association: 41–49.

Couper, M. P. & Groves, R. M. (1992). Interviewer reactions to alternative hardware for computer-assisted personal interviewing. *Journal of Official Statistics*, *8(2)*, 201–210.

Couper, M. P. & Lyberg, L. (2005). The use of paradata in survey research. Proceedings of the International Statistical Institute Meetings.

Couper, M. P. & Wagner, J. (2011). Using paradata and responsive design to manage survey nonresponse. ISI World Statistics Congress, Dublin, Ireland.

Dekker, K., English, N., Winfrey, K., & Seeger, J. (2013). Cost implications of new address listing technology: efficiency and data quality. Federal CASIC Workshops. Washington, DC.

Dielman, L. & Couper, M. P. (1995). Data quality in a CAPI survey: keying errors. *Journal of Official Statistics*, *11(2)*, 141–146.

Drew, J. H. & Fuller, W. A. (1980). Modeling nonresponse in surveys with callbacks. Proceedings of the Section on Survey Research Methods of the American Statistical Association.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174(4)*, 1029–1049.

Dutwin, D., Loft, J. D., Darling, J. E., Holbrook, A. L., Johnson, T. P., Langley, R. E., … Zukerberg, A. (2015). Current knowledge and considerations regarding survey refusals: executive summary of the AAPOR task force report on survey refusals. *Public Opinion Quarterly*, *79(2)*, 411–419. doi:doi:10.1093/poq/nfv025

Eckman, S., Sinibaldi, J., & Mantmann-Hertz, A. (2013). Can interviewers effectively rate the likelihood of cases to cooperate. *Public Opinion Quarterly*, *77(2)*, 561–573.

Ellis, C., Sikes, N., Sage, A., Eyerman, J., & Burke, B. (2011). Technological advances to reduce survey error. Paper Presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Fowler Jr., F. J. & Mangione, T. W. (1990). *Standardized survey interviewing: minimizing interviewer-related error*. Newbury Park, California: Sage.

Fowler, F. J. (1991). Reducing interviewer-related error through interviewer training, supervision, and other means. In P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 259–278). New York: John Wiley & Sons.

Goodchild, M. F., Nusser, S. M., Pickle, L. W., Lane, J., Williamson, P., Mulry, M. H., … Hox, J. (2007). The Morris Hansen Lecture 2006 statistical perspectives on spatial social science. *Journal of Official Statistics*, *23(3)*, 1–15.

Groves, R. M., Cantor, D., Couper, M. P., Levin, K., McGonagle, K., Singer, E., & Van Hoewyk, J. (1997). Research investigations in gaining participation from sample firms in the current employment statistics program. Proceedings of the Section on Survey Research Methods.

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M. & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169(3)*, 439–457.

Groves, R. M. & McGonagle, K. A. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17(2)*, 249–266.

Haddaway, S. (2013). CAPI surveys on android devices in the developing world. Federal CASIC Workshops. Washington, DC.

Kalsbeek, W. D., Botman, S. L., Massey, J. T., & Liu, P.-W. (1994). Cost-efficiency and the number of allowable call attempts in the National Health Interview Survey. *Journal of Official Statistics*, *10(2)*, 133–152.

Kalsbeek, W. D., Mendoza, O. M., & Budescu, D. V. (1983). Cost models for optimum allocation in multi-stage sampling. *Survey Methodology*, *9(2)*, 154–177.

Keating, M., Loftis, C., McMichael, J., & Ridenhour, J. (2014). New dimensions of mobile data quality. Federal CASIC Workshops. Washington, DC.

Kirgis, N. & Lepkowski, J. (2013). Design and management strategies for paradata-driven responsive design: illustrations from the 2006-2010 National Survey of Family Growth. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 121–144). Hoboken, NJ: Wiley.

Krenn, P. J., Titze, S., Oja, P., Jones, A., & Ogilvie, D. (2011). Use of global positioning systems to study physical activity and the environment: a systematic review. *American Journal of Preventive Medicine*, *41(5)*, 508–515.

Kreuter, F. (2013). *Improving surveys with paradata: analytic use of process information*. Hoboken, NJ: Wiley.

Kreuter, F. & Kohler, U. (2009). Analyzing contact sequences in call record data. potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, *25(2)*, 203–226.

Kreuter, F., Mercer, A., & Hicks, W. (2014). Increasing fieldwork efficiency through prespecified appointments. *Journal of Survey Statistics and Methodology*, *2(2)*, 210–223.

Kulka, R. A. & Weeks, M. F. (1988). Toward the development of optimal calling protocols for telephone surveys: a conditional probabilities approach. *Journal of Official Statistics*, *4(4)*, 319–332.

Kurkowski, K. (2013). Using internet and hand-held computers for data collection in poland. Paper presented at The 5th International Workshop on Internet Survey and Survey Methodology. Daejon, Republic of Korea.

Lemmens, M. (2011). *Geo-information: technologies, applications and the environment*. Dordrecht: Springer Science+Business Media.

Levinsohn, D. D., J. R. aand Medeiros, Duke, J. C., Yost, P. A., Litavecz, S. D., Zhang, Y., & Karlsen, R. (2010).

The General Survey System (GSS): a mobile technologies system for collecting and managing study data. The 138th Annual Meeting of the American Public Health Association. Denver, CO.

Luiten, A. & Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: an experiment in the survey of consumer satisfaction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176(1)*, 169–189.

Mavoa, S., Oliver, M., Witten, K., & Badland, H. M. (2011). Linking GPS and travel diary data using sequence alignment in a study of children's independent mobility. *International Journal of Health Geographics*, *10(1)*, 1–10.

Miller, P. (2014). What does adaptive design mean to you? Federal CASIC Workshops. Washington, DC.

Morton, K. B., McMichael, J. P., Cajka, J. C., Curry, R. J., Iannacchione, V. G., & Cunningham, D. B. (2007). Linking mailing addresses to a household sampling frame based on census geography. JSM Proceedings.

Nusser, S. M. (2007). Using geospatial information resources in sample surveys. *Journal of Official Statistics*, *23(3)*, 285–289.

Nusser, S. M. & Fox, J. E. (2002). Using digital maps and GPS for planning and navigation in field surveys. Iowa State University Department of Statistics: Iowa State University.

Olson, K. & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 43–72). Hoboken, NJ: Wiley.

Olson, K. & Wagner, J. (2015). A field experiment using GPS devices to monitor interviewer travel behavior. *Survey Research Methods*, *9(1)*, 1–13.

Ongena, Y. P. & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics*, *22(3)*, 419–451.

Peytchev, A., Baxter, R. K., & Carley-Baxter, L. R. (2009). Not all survey effort is equal: reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, *73(4)*, 785–806.

Peytchev, S., A. Riley, Rosen, J., Murphy, J., & Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, *4(1)*, 21–29.

Piras, M. & Cina, A. (2010). Indoor positioning using low cost GPS receivers: tests and statistical analyses. 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN), IEEE.

Pline, J. L. (Ed.). (1992). *Traffic engineering handbook*. Washington, DC: Prentice-Hall.

Potthoff, R. F., Manton, K. G., & Woodbury, M. A. (1993). Correcting for nonavailability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, *88(424)*, 1197–1207.

Presser, S. & McCulloch, S. (2011). The growth of survey research in the united states: government-sponsored surveys, 1984-2004. *Social Science Research*, *40(4)*, 1019–1024.

Rempel, R. S. & Rodgers, A. R. (1997). Effects of differential correction on accuracy of a GPS animal location system. *The Journal of Wildlife Management*, *61(2)*, 525–530.

Seeger, J. (2011). A mobile, GPS-enabled listing application. Federal CASIC Workshops. Washington, DC.

Sikes, N. (2009). Current trends in mobile technology for survey research. Federal CASIC Workshops. Washington, DC.

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, *16(3)*, 350–369.

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

U.S. Census Bureau. (2014). GPS coordinates. Retrieved from http://www.census.gov/about/policies/privacy/data%5C_protection/gps%5C_coordinates.html

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, *7(1)*, 45–55.

Wagner, J. & Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, *2(3)*, 323–342.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, *28(4)*, 477–499.

Walejko, G. & Wagner, J. (2015). Challenges to innovation in face-to-face surveys posed by interviewer noncompliance. Paper presented at the Annual Conference of the American Association for Public Opinion Research.

Weeks, M. F., Jones, B. L., Folsom Jr., R. E., & Benrud, C. H. (1980). Optimal times to contact sample households. *Public Opinion Quarterly*, *44(1)*, 101–114.

Weeks, M. F., Kulka, R. A., & Pierson, S. A. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, *51(4)*, 540–549.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176(1)*, 211–225.

Wood, A. M., White, I. R., & Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169(3)*, 525–542.

Wu, J., Jiang, C., Liu, Z., Houston, D., Jaimes, G., & McConnell, R. (2010). Performances of different global positioning system devices for time-location tracking in air pollution epidemiological studies. *Environmental Health Insights*, *4*, 93–108.

Yan, T. & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 73–96). Hoboken, NJ: Wiley.

Zandbergen, P. A. (2009). Accuracy of iphone locations: a comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*, *13(1)*, 5–26.

Appendix

Additional Details on Sampling and Linking

In order to draw a sample, we needed decide what would be the appropriate unit for the sampling. We could have sampled PSUs, interviewers, interviewer-days (where each day that any interviewer worked is a unit) or GPS points within a defined area. We sampled a subset of interviewer-days because this is the unit of analysis for which the GPS points were collected, making the interviewer-day a better unit to sample than PSUs, interviewers, or GPS points within a defined area. This sampling design provided some protection against missing true linkages due to errors in the recorded times in each source of data. Errors in the dates in the GPS data and call records are less likely to occur than errors in the time of the attempt. The interviewer survey indicated that most call records are generated on the day the call attempt was made (79% said they always or almost always generate the call record immediately after the attempt, 92% said they never complete call records at the "end of the week" in which the call attempt was made). This parallels the experience reported by the Census Bureau (Bates et al., 2010). The location of the roads (streets, highways, etc.) was determined using the Census TIGER files. The snapping of the GPS points gathered by the smartphones to these roads was done using ArcGIS and the Geospatial Modelling Environment (GME). Then routes were created by using the ArcGIS Network Analysis toolbox. Merging the call records to the routes then involved several steps. For each interviewer and interviewer-day, we calculated the distance between the sampled housing unit and the inferred route, the difference between the time of the route and the time of the call attempt, and identified whether the sampled housing unit was on the same street as the inferred route (in order to find the road that the housing unit faces). Sampled and visited housing units were eligible to be linked if

1. the distance between sampled housing unit and the route is within 1,000 meters,

2. the street name of sampled housing units' address and street name of the route match, and

3. the route is one of three closest routes to the sampled housing unit.

The latter stipulation was necessary to limit the computation required in situations where many routes would be within 1,000 meters of the sampled housing unit.