DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 211–237, http://dx.doi.org/10.1515/JOS-2018-0010

# An Analysis of Interviewer Travel and Field Outcomes in Two Field Surveys

*James Wagner*[1] *and Kristen Olson*[2]

In this article, we investigate the relationship between interviewer travel behavior and field outcomes, such as contact rates, response rates, and contact attempts in two studies, the National Survey of Family Growth and the Health and Retirement Study. Using call record paradata that have been aggregated to interviewer-day levels, we examine two important cost drivers as measures of interviewer travel behavior: the distance that interviewers travel to segments and the number of segments visited on an interviewer-day. We explore several predictors of these measures of travel – the geographic size of the sampled areas, measures of urbanicity, and other sample and interviewer characteristics. We also explore the relationship between travel and field outcomes, such as the number of contact attempts made and response rates. We find that the number of segments that are visited on each interviewer-day has a strong association with field outcomes, but the number of miles travelled does not. These findings suggest that survey organizations should routinely monitor the number of segments that interviewers visit, and that more direct measurement of interviewer travel behavior is needed.

*Key words:* Interviewer travel; survey costs; nonresponse; paradata.

## 1. Introduction

In face-to-face surveys, survey organizations use multi-stage area probability samples with clustering of sampled housing units (i.e., 'area segments,' Kish 1965) in order to constrain travel costs, but we know very little about interviewer travel behavior in these surveys. In general, an interviewer travels from his/her home to a sampled neighborhood. Once in the sampled neighborhood, the interviewer makes contact attempts to several sampled housing units, identifying potentially eligible respondents ("screening") and conducting interviews.

[1] University of Michigan, Institute for Social Research, 426 Thompson St. Room 4050, Ann Arbor, MI 48104, U.S.A. Email: jameswag@umich.edu
[2] University of Nebraska-Lincoln, Department of Sociology, 703 Oldfather Hall, Lincoln, NE 68588, U.S.A. Email: kolson5@unl.edu
**Acknowledgments:** An earlier version of this article was presented at the Joint Statistical Meetings, August 2011, Miami Beach, FL. The authors thank the NSFG and the HRS projects for access to these data, and Brady West, four anonymous reviewers and the Associate Editor for comments on an earlier draft. In particular, the Associate Editor gave us the insight that traveling longer distances at higher speeds (e.g., on a freeway) might require less time than a shorter distance at a slower speed (e.g., city street). The NSFG 2006–2010 was conducted by the Centers for Disease Control and Prevention's CDC's National Center for Health Statistics (NCHS), under contract #200-2000-07001 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or the other funding agencies. The HRS is funded by the National Institute on Aging (U01 AG009740), with supplemental support from the Social Security Administration.

If a sampled area is quite far from the interviewer's home, large amounts of time are spent 'on the clock' for travel, not screening, interviewing, or otherwise recruiting sampled units. As a result, interviewer travel is a large component of the total budget in personal visit surveys; for many large national surveys, travel costs are 20% to 40% of the total survey budget (e.g., Judkins et al. 1990; Kalsbeek et al. 1994; Weeks et al. 1983; Sudman 1965–66; Sudman 1967). Interviewers may vary in their ability to plan efficient travel which may impact their efficiency and other field outcomes. Yet surprisingly unexplored are characteristics of interviewer travel behavior – measured either through the actual number of miles traveled or the number of area segments visited on a given interviewer workday – and the association between interviewer travel behavior and field outcomes, such as the number of contact attempts, contact and cooperation rates. These interviewer travel behaviors and field outcomes are important drivers of both response rates and costs.

In this article, we examine this important but little explored component of field surveys – interviewer travel behavior – in two large national surveys in the United States. We also investigate the relationship between the distance that interviewers travel while performing their work and field outcomes – that is, the cumulative results of an interviewer's actions when visiting sampled housing units in the field. In particular, we examine the following two questions:

1. What are predictors of interviewer travel behavior in two large national US field surveys?
2. Is interviewer travel behavior associated with field outcomes (i.e., number of contact attempts, contact rates, screening, and main interview cooperation rates)?

This article takes a first observational look at the relationship between characteristics of areas, travel-related costs, and survey errors. We will examine these relationships using cross-classified random effects models to analyze interviewer travel and field outcomes as indicated through call history and timesheet data from two large-scale national area probability sample surveys in the United States – the National Survey of Family Growth and the Health and Retirement Study. Specifically, for our first research question, we examine predictors of interviewer travel such as the geographic size of sampled areas and urbanicity. For our second research question, we examine whether interviewer travel decisions contribute to survey errors, including variability in nonresponse rates across interviewers and decisions that may lengthen the field period. We anticipate that interviewers who travel more miles will also have more contact attempts because the travel allows them to make these contact attempts, but lower contact and response rates because the time spent traveling constrains the amount of time that can be spent administering a questionnaire. Understanding the relationship between travel and field outcomes will aid survey practitioners in designing efficient surveys.

## 2. Background

### *2.1. Interviewer Travel Overview*

Although all field studies require interviewers to travel to sampled housing units, there is surprisingly little empirical examination of interviewer travel behavior. Travel in sample surveys has been specified as a constraint on sampling error, such as the variance of an estimated mean from a cluster sample (Sudman 1978; Hansen et al. 1953, 274). There are different measures of interviewer travel behavior available from paradata, including how many segments an interviewer visits on a given day or the distance traveled on these trips. Paradata include call record data (sometimes called contact history data) and timesheet and travel expense data. Cost models indicate three inputs that are needed – (1) the number of times a segment is visited, (2) the distance between the interviewer's home and the segments, and (3) the distance traveled within segments (Sudman 1967, ch. 2; Judkins et al. 1990). For example, Cochran (1977, 183) provides a cost model for a cluster sample with cost of travel between clusters as $C = c_1 n + c_t \sqrt{n} + c_2 nm$, where $n$ is the number of clusters, $m$ is the number of units within each cluster, $c_1$ is the cost of measuring (an unspecified combination of listing, screening, and interviewing) a cluster, $c_t$ is the cost related to travel between clusters, and $c_2$ is the cost of travel within a cluster (see also Kalsbeek et al. 1983; Judkins et al. 1990). Despite this central role of cost models for determining optimal sample designs, empirical evaluation of what these cost inputs are in actual field studies is lacking.

Existing cost data are largely for one survey (the National Household Interview Survey, or NHIS, conducted for the US National Center for Health Statistics). Kalsbeek et al. (1983) note that empirical data are not available and, therefore, attempt to use simple geometry to model travel costs. In examining cost drivers in the 1988 NHIS, Judkins and colleagues (1990) note that there is only limited empirical data available on the number of segments visited and on distance traveled. They find that the average workload size for interviewers was 2.3 segments, but were not able to empirically evaluate mileage traveled. Chen (2012) draws on data about field outcomes from the 2004 NHIS in order to simulate travel behaviors. Thus, there is a very limited literature with almost no empirical data on travel costs as they relate to sampling error. There is even less information about how these travel behaviors are related to field outcomes.

Further, the manner in which interviewers make decisions about travel is not well understood. Little information on interviewer travel can be obtained from training materials, as interviewers receive limited training on travel; the training that they do receive generally focuses on cost containment. For example, interviewers are trained to monitor (Mayer 1968) or limit the number of trips to sampled segments to keep costs under control (Morton-Williams 1993, 141; Campanelli et al. 1997, 3–20). They are also trained to make contact attempts at times during which they are likely to reach someone at home, varying the day and time of contact attempts (e.g., Morton-Williams 1993; Stoop et al. 2010), times which may vary across sample units (e.g., Durrant and Steele 2009; Blom 2012). In practice, interviewers report visiting multiple cases once they have traveled to a particular sampled area (Peachman 1992). Interviewers may then travel between segments to reach different sampled persons at home, even revisiting area segments on the same day. The extent to which they do this is unknown.

One potential reason that interviewer travel has received limited attention is that studying interviewer travel outcomes requires a record to be kept of travel itself. In most in-person surveys, this can be obtained in three ways: (1) mileage reports from the interviewer when asking for travel reimbursement on their timesheets, (2) distances obtained from geocoding the locations of the sampled housing units as recorded in the call records, or (3) summaries of the number of sampled neighborhoods (segments) visited by each interviewer obtained through aggregating the call records or reports by the interviewer themselves. All three sources are likely to contain errors. Nevertheless, evaluating existing travel information is necessary to establish that there are associations between travel behaviors and field outcomes such as contact and cooperation rates, even with imperfect data. If interviewers are unsuccessful at establishing contact or completing interviews, then they will be more likely to spend time in travel as they continue to seek contact with a selected housing unit. In this article, we will focus on mileage reports and a summary of the number of segments visited on a given interviewer-day obtained through the call records.

### 2.2. *Predictors of Interviewer Travel Behavior*

The first research question addresses predictors of interviewer travel behavior. Interviewers are not randomly assigned to segments, and their skill sets with respect to travel decisions vary. We expect that travel will differ overall for different geographic areas, interviewers with different levels of experience, and at different times during the field period.

Given higher population density in urban areas, *we anticipate that interviewers will have to travel more miles in rural areas than in suburban or urban areas because the segments are larger in rural areas*. Thus, we include urbanicity in all of the models. *We expect that larger areas and areas considered to be "more difficult" will require more travel*. Therefore, we include the area (measured in square miles) of the primary sampling units to account for the variation in area between Primary Sampling Units (PSUs). In order to account for variation between PSUs in the difficulty of obtaining an interview, we use the United States Census Bureau's "Hard-to-Count" score. This score is based upon characteristics of Census Tracts, including demographic characteristics predictive of response rates to the mailed portion of the Decennial Census (Bruce and Robinson 2006). We have created a weighted average, using the Census Tract count of housing units as the weight of this score for each PSU in the sample. We also anticipate that interviewers with less experience will travel differently than experienced interviewers (as interviewers with different experience levels have different calling times, Campanelli et al. 1997). In particular, *we expect that inexperienced interviewers will be more likely to follow travel suggestions received in training or from supervisors than experienced interviewers*. As such, in surveys where interviewers are instructed to visit all active, sampled housing units in a particular segment, *we expect inexperienced interviewers will be more likely to visit all active, sampled housing units, and thus do more within-segment traveling*. As such, we account for interviewer experience in all of the models as well.

Finally, we expect differences in travel overall for different times in the field period. In surveys where low yield segments are removed (either via two-phase sampling or by a management decision) from the active sample, *we expect that interviewers will tend to travel less at later times in the field period.*

### 2.3. *Association Between Interviewer Travel Behavior and Field Outcomes*

Decisions about travel affect not only costs, but also are likely to be associated with field outcomes. This may be either because of interviewers traveling in order to make more call attempts when early attempts do not yield interviews or because the amount of time spent traveling constrains the time possible to conduct an interview. Thus, our second research question addresses whether interviewer travel behavior is associated with field outcomes such as the number of contact attempts, contact rates, and response rates.

We are aware of only two simulation studies that address this question. One of these simulation studies found that smaller PSU sizes give interviewers more time to contact households because of reduced travel time (Chen 2012). The other simulation study suggested that the length of interview may modify the effect that travel has on field outcomes in that shorter interviews yield more time for travel and increase the possibility of obtaining additional interviews (Bienias et al. 1990). Although interviewers are a key source of variability in response, contact and cooperation rates in face-to-face surveys (e.g., O'Muircheartaigh and Campanelli 1999; Campanelli et al. 1997; Purdon et al. 1999; Pickery and Loosveldt 2002; Durrant and Steele 2009; Blom et al. 2011; Blom 2012; Stoop et al. 2010), to our knowledge, no previous study has examined the association between interviewer travel behaviors and contact and cooperation rates using actual data.

The two previous simulation studies of interviewer travel (e.g., Bienias et al. 1990; Chen 2012) suggest *two competing hypotheses* about the relationship between travel behaviors and field outcomes. Both hypotheses assume that interviewers make their own decisions about which sampled segments to visit and which housing units to attempt on a given interviewer-day, that interviewers no longer visit sampled cases once an interview is completed, and that there are no constraints put on interviewers for the number of attempts that they can make on a given day or to a sampled housing unit. The first hypothesis posits a positive association between travel and the number of contact attempts made on a given day – interviewers who drive more miles or visit more segments are traveling to make more contact attempts. The second hypothesis anticipates a negative association between travel and contact attempts, arguing that interviewers who spend more time travelling – and thus have more miles driven or more segments visited – have less 'on the ground' time to make contact attempts for a fixed amount of work time in a given day.

We anticipate that the association between travel and the number of contact attempts will differ for different types of field outcomes. With the same assumptions as above, in particular, *we expect a positive association between travel behaviors and the number of contact attempts made on an interviewer day*. That is, interviewers who travel more will be doing so in order to make additional contact attempts. *We anticipate, however, that travel behaviors will be negatively associated with screening and response rates*. Interviewers who make contact with a household or obtain interviews will have less time to travel to other areas (that is, more successful field outcomes lead to less travel).

Because interviewer training on travel is generally linked to minimizing the number of trips made to individual segments (sampled neighborhoods), *we anticipate that the total number of segments visited will be more likely to be associated with field outcomes than the actual number of miles traveled*. Yet costs are directly related to the total number of miles that an interviewer drives. We thus will examine both measures of interviewer travel.

## 3. Data

### 3.1. Surveys

We examine two large-scale, national face-to-face surveys conducted by the Survey Research Center at the University of Michigan – the National Survey of Family Growth (NSFG) and the Health and Retirement Study (HRS). The two surveys differ in scope, target populations, and field periods. Additionally, the NSFG is a cross-sectional survey and the HRS is a longitudinal survey.

### 3.2. National Survey of Family Growth

The NSFG 2006–2010 Continuous was carried out under a contract with the US Center for Disease Control and Prevention's (CDC) National Center for Health Statistics. The NSFG collects information about fertility, childbearing, and sexual behaviors among women and men in the US aged 15 to 44. Thus, the NSFG interview process had two steps – identifying an age-eligible respondent through screening and conducting a "main" survey interview. NSFG 2006–2010 continuous released fresh samples quarterly, with a twelve-week field period, in what is called a "continuous sample design" (Lepkowski et al. 2010). Interviewers are generally assigned three neighborhoods or segments within a single, "home" PSU per quarter. The twelve-week field period is divided into two phases. At the end of ten weeks, a subsample of two segments (with additional subsampling of lines within those two segments) is selected for each interviewer. The data analyzed for this article come from the second quarter which ran from September to December of 2006. The initial sample included 5,063 housing units.

The assignment of interviewers to sampled segments is not random; in most PSUs there is only a single interviewer who is assigned a random sample of segments from within the PSU. When there is more than one interviewer, the interviewers are assigned segments near their home location to try to minimize travel. Overall, 45% of interviewers visited only one PSU over the field period, and 55% visited two or more PSUs. Additionally, 29% of PSUs had only one interviewer visit them, with the remaining 71% having at least two different interviewers visit them. Thus, interviewer-days are cross-classified between interviewers and PSUs. In some PSUs, including all of Alaska and Hawaii, interviewers are flown in for field work and do not travel from their home location. As such, we will exclude Alaska and Hawaii from these analyses.

The number of days worked by each of the 41 interviewers ranges from 2 (one interviewer) to 72, with only three interviewers working fewer than ten days, and an average of 44.6 days. On 97% of the days, interviewers remained within a single PSU.

All of the NSFG field staff record call records and timesheets electronically. Information about the housing units each interviewer visits and the corresponding field outcomes are recorded in call records. Call record information is usually filled out daily or more frequently (after each contact attempt). Travel information is recorded by the interviewer in their timesheets for purposes of mileage reimbursement in personal vehicles, and is submitted at most daily (the interviewers are paid biweekly, so payment information is not always submitted on the day that it occurred). The two systems, however, do not interface directly. That is, interviewers are not probed in their timesheets to record time specifically related to effort recorded in the call records. As a result, there can be a "mismatch" between activities as recorded in the call records and activities as recorded in timesheets and expense reports. We will examine information from both the call records (aggregated to the number of segments visited per day) and the timesheets (reported mileage per day) in this analysis. For the data used in our models, each record represents a summary of the travel and field outcomes for a day that an interviewer worked.

### *3.3. Health and Retirement Study*

The Health and Retirement Study (HRS) is a panel study of U.S. adults aged 50 and above, with initial data collection starting in 1992. Every six years, a new age-eligible sample is recruited to include the newly aged-in 51 to 56 year olds. In this article, we examine the new cohort added during the 2004 data collection. The HRS uses an area probability sample to identify households with newly age-eligible adults (Heeringa and Connor 1995; Health and Retirement Study 2008), with two interviewing steps (screening households for age-eligible respondents and "main" interviewing) for the new cohort. The sample in all PSUs and segments (40,120 housing units) was released at the beginning of the field period. Unlike the NSFG, the HRS did not select a second phase sample of segments. Near the end of the field period, with only 790 sample housing units still active, a subsample of half of these housing units was selected.

The panel and the new cohort of the HRS are recruited using somewhat overlapping field staff at the beginning of data collection. We examine here the effort of the interviewers who were mainly assigned to screen households and identify persons who were eligible for the new cohort and to interview them. Among these interviewers, 91.7% of trips had no contact attempts to the panel cases sample, and 5.2% had only one contact attempt to the panel cases. As such, although the HRS is a longitudinal study, the data that we examine here are analogous to a cross-sectional study.

In the HRS, multiple interviewers are (not randomly) assigned to a PSU, and each interviewer was assigned to a variable number of secondary sampling units (or segments) based on their geographic proximity to the segments, again having a cross-classified data structure for interviewer-days, nested across both interviewers and PSUs. In general, HRS field managers had more flexibility in assigning segments to interviewers than managers for the NSFG. For the HRS, there are generally more interviewers in each PSU. Segments could be assigned to the interviewer who lived close to them. Overall, each PSU had at least two interviewers work in it, ranging from two to 35 different interviewers in a single PSU. Additionally, 45% of interviewers worked in only one PSU, but the remaining 55% worked in two or more different PSUs. The field period for the HRS was about twelve months. Further, instead of having new segments assigned each quarter, HRS interviewers had all of their segments available from the beginning of the field period. There was much greater variability in the number of days worked by HRS interviewers compared to the NSFG interviewers, ranging from one day to 255 days, with an average of 48.8 days. To increase the stability of estimates (Olson and Peytchev 2007), only interviewers who had worked at least ten days were included in the analyses. This excluded 61 interviewers who had collectively worked a total of 176 days, 37 of whom had worked only one day during the field period, with 205 interviewers remaining who had collectively worked 12,940 days, for an average of 62.2 interviewer-days per interviewer.

As with the NSFG, the data of interest for the HRS also come from timesheet information and from the call records. Both timesheets and call records are maintained electronically, and call records are filled out daily or more often. As in the NSFG, interviewers are paid biweekly. As with the NSFG, information about the number of miles traveled comes from the timesheet information; field outcomes and the number of segments visited on each interviewer day come from the call records. The models for the HRS are estimated on data for which each record represents a summary of the travel and field outcomes for a day worked by an interviewer.

## 4. Methods

### *4.1. Travel Variables*

Ideally, the level of analysis for each interviewer and sampled unit for this study would be at the call attempt (sometimes called contact attempt) level. Each sampled address visited during a given trip to a sampled segment could be geocoded and the distance between the interviewer's house, the first sampled address, and among sampled addresses could be calculated. However, the timesheet data are kept at the day level for each interviewer, not at the contact attempt level. Thus, we aggregate information for each interviewer for each day of the survey period, and analyze travel at an interviewer-day level.

We use two measures of interviewer travel behavior: (1) the total number of miles reported per day in the timesheet and (2) the total number of trips made to sampled segments visited during an interviewer-day as calculated from the call records. First, we use the total number of miles traveled on a given day that the interviewer reported in their timesheet for purposes of reimbursement. In the NSFG, over 90% of days (90.9%) have mileage data reported, and in the HRS, 96.2% of days have mileage data reported. Among the days with mileage data reported, interviewers travel an average of 85.4 miles per interviewer-day in the NSFG and 53.4 miles per interviewer day in the HRS.

Second, to obtain a count of the number of trips made to segments during a day, we define trips as any visit to a segment that involves travel between the interviewer's home and the segment, or between segments. For example, if an interviewer travels to segment A from their home, then goes to segment B, and then returns to segment A, they have taken three trips to sampled segments. For parsimony, we will refer to this as a "three segment" trip. Interviewers in the NSFG and HRS visited a similar number of segments per interviewer day – 1.85 segments per interviewer-day in the NSFG and 1.91 segments per interviewer-day in the HRS. The distribution of number of segments visited per interviewer-day is remarkably similar across the two surveys. In both surveys, roughly 54% of all interviewer-days are spent visiting one segment, and just fewer than 22% of interviewer-days involve contact attempts to two segments. The balance of interviewer-days (24%) had contact attempts in more than two segments. In the NSFG, the correlation between the number of segments visited and the total number of miles traveled is 0.23 ($p < .0001$), whereas it is only 0.06 ($p < .0001$) in the HRS. These differences likely reflect the greater flexibility that HRS has in assigning segments to interviewers.

One possible limitation is the quality of the travel information. Discrepancies between the number of segments calculated from the call records and miles reported by the interviewer on a timesheet can occur for a number of reasons. For example, interviewers can record work-related travel – such as travel to a training session – that is not related to field effort, interviewers who are flown into a sampled PSU will not record mileage because they do not need to be reimbursed for rental car mileage, and interviewers who work in major metropolitan areas may not use a car to travel among sampled units, instead using public transportation. Additionally, interviewers may fail to complete a call record for some types of travel, such as driving by a house and not seeing evidence of anyone at home, or may not complete their travel reports until the end of the day, potentially forgetting a trip or misremembering where they traveled. Interviewers may also enter travel reimbursement information for the wrong date (Wang and Biemer 2010; Biemer et al. 2013; Wagner et al. 2013). Another type of error occurs when the wrong mode for a contact attempts is entered into the record. If a telephone visit is entered into the records as a face-to-face call, then incorrect assumptions about travel will be made.

To address the potential limitation of the quality of the mileage reported in the timesheets, for the NSFG, we also conducted a sensitivity analysis. We used information about the interviewer's home address to measure distance traveled on a given day starting from an interviewer's house via geocoded addresses obtained in the call records. We geocoded each interviewer's home address and the centroid of the sampled segment, and calculated the distance in miles 'as the crow flies' from the interviewer's home address to their segments and among their segments. Although other options are available for calculating distance (such as 'best routes' calculated through Google Maps), we started with this approach for simplicity. Given missing data on interviewer home addresses and limitations of the geocoding software, we were unable to geocode addresses for 24.6% of interviewer-days in the NSFG. The two sources of travel data are highly, but not perfectly correlated ($r = 0.780$, $p < .0001$). In general, the interviewer-reported timesheet data tends to be higher than the geocoded data – as would be expected given our use of the 'as the crow flies' distance for the geocoded data. We conducted all of our analyses using both metrics and came to identical conclusions. In the HRS, we do not have the interviewer's home address and thus cannot calculate a geocoded distance from the interviewer's home.

### *4.2. Field Outcome Variables*

We examine five field outcomes (see Table 1) as the dependent variables in our models:

1. Total number of contact attempts made to sample housing units on a given interviewer-day for screening cases,
2. total number of contact attempts made for main interviews,
3. contact rates,
4. screener interview rates, and
5. main interview rates.

In both surveys, we separate contact attempts into two groups – contact attempts to screen the household for an eligible sample person and contact attempts to complete the main

*Table 1. Descriptive Statistics of Travel, Field Outcomes, Urbanicity and Experience across Interviewer-Days, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).*

| Variable | NSFG Mean (SD) or % | HRS Mean (SD) or % |
|---|---|---|
| Travel | | |
| Number of miles | 85.42 (74.55) | 53.36 (50.08) |
| Number of segments | 1.85 (1.18) | 1.91 (1.22) |
| 1 segment | 54.9% | 53.5% |
| 2 segments | 21.9% | 21.5% |
| 3 segments | 12.2% | 12.3% |
| 4 segments | 5.0% | 6.1% |
| 5+ segments | 6.1% | 6.7% |
| Day-level Field Outcomes | | |
| Number of screening contact attempts | 10.68 (11.58) | 11.02 (11.25) |
| Number of main interview contact attempts | 4.08 (4.03) | 1.43 (2.08) |
| Contact rates | 0.43 (0.30) | 0.36 (0.29) |
| Screener interview rates | 0.21 (0.23) | 0.22 (0.25) |
| Main interview rates | 0.30 (0.34) | 0.20 (0.36) |
| Urbanicity | | |
| Largest MSAs | 17.2% | – |
| Smaller MSAs | 41.2% | – |
| Non-MSAs | 41.6% | – |
| Self-representing PSUs | – | 34.4% |
| Non self-representing PSUs | – | 65.6% |
| Interviewer Experience | | |
| No experience | 9.4% | 23.9% |
| Any prior experience | 90.6% | 76.1% |
| Number of weeks in field period | 6.51 (3.35) | 28.83 (11.91) |
| Census Hard-to-Count Score | 34.22 (12.16) | 40.67 (19.40) |
| Area (Square miles)* | 2,200 (1,464) | 2,200 (3,378) |

*Note:* Number of interviewer-days = 1,784 in the NSFG and 12,940 in the HRS. SD = Standard Deviation; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.
*Rounded to nearest hundreds.

interview. For both of these surveys, screening and main interviews are two separate activities that usually occur on different days (75.7% and 90.7% of completed screening interviews for the NSFG and HRS respectively required additional main contact attempts on another day). This is because the person completing the screening interview may be different than the person selected to complete the main interview and because the main interviews for both surveys take a relatively long time to complete. Therefore, interviews need to be scheduled at times that are convenient for the sampled person. In both surveys, there are fewer main interview attempts than screener interview attempts due to the length of the main interview and the eligibility criteria that had to be met to conduct a main interview. In the NSFG, there is an average of 10.68 screener attempts and 4.08 main interview attempts per interviewer-day. In the HRS, there is an average of 11.02 screener attempts and 1.43 main interview attempts per interviewer-day. While the average number of screener attempts is very similar across the two surveys, the number of main attempts is much higher for the NSFG. This is likely due to the higher eligibility rate for the NSFG compared to the HRS, and that the HRS often attempts to interview two members of the household on the same day. The same number of screener attempts will produce more main attempts. Since these are averages across days, it also indicates that NSFG interviewers may have worked longer shifts.

The contact, screener interview, and main interview rates are calculated at an interviewer-day level. In particular, the contact rate is the total number of attempts with contact divided by the total number of attempts made on a given interviewer-day. The screener interview rates are the total number of completed screeners divided by the total number of attempts for screener interviews (those to cases of previously unknown eligibility), and the main interview rates are similarly the total number of completed main interviews divided by the total number of attempts for main interviews (those to cases of known eligibility). In the NSFG, the average contact rate across all interviewer-days is 43.2%, and the average contact rate for the HRS is 35.5%. Over all of the interviewer days, in the NSFG, the average screener interview rate is 21.3%, and the main interview rate, conditional on known eligibility is 29.5%. In the HRS, the screener rate is 21.8%, and the main interview rate among known eligible persons is 20.4%. These call-level contact, screener and main interview rates are different from the final case-level contact rate, screener and main interview rates for the survey which are calculated by the total number of cases contacted or completed by the end of the study divided by the total number of sampled (eligible) cases.

### *4.3. Predictor Variables for Multivariate Models*

The predictors are chosen from a set that have been shown or are hypothesized to be related to interviewer travel and productivity. Since travel may be impacted by characteristics of the interviewer, the PSU, and characteristics of the available sample, we select predictors from our model from each of these areas. We model the number of segments visited and miles travelled each day, including as predictors in the models urbanicity, interviewer experience, a continuous measure of the week in the field period, a measure of how difficult the area is to enumerate based on the 2000 Decennial Census Hard-to-Count (HTC) score (Bruce and Robinson 2006), and the area of the PSU in square miles. Urbanicity is measured from characteristics of the PSU in which an interviewer's segments are located. In the NSFG, urbanicity has three levels – the largest Metropolitan Statistical Areas (MSAs, 17.2% of interviewer-days), smaller MSAs (41.2%) and non-MSAs (41.6%). A Metropolitan Statistical Area is an urban geographic area with at least 50,000 residents (United States Census Bureau 2013). In the HRS, we use an indicator for whether the PSU was self-representing or not; 34.4% of interviewer-days occurred in self-representing PSUs. In both surveys, interviewers were quite experienced – 90.6% of NSFG interviewers and 76.1% of HRS interviewers had prior interviewing experience. The week of the field period ranged from 1 to 12 for the NSFG and from 1 to 52 for the HRS. We include an indicator for whether the interviewer-day was in Phase $1(=1)$ versus Phase $2(=0)$, and an interaction term between Phase 1 and week in the field period to account for potential nonlinearities during "close-out" periods of each survey. In the NSFG, Phase 1 was defined as all weeks before the 11th week, and in the HRS, Phase 1 was defined as all weeks before the 48th week. The mean Census hard-to-count score was 34.22 (SD = 12.16) in the NSFG and 40.67 (SD = 19.40) in the HRS. Finally, the average PSU was about 2,200 square miles (SD = 1,464) in the NSFG and also about 2,200 square miles (SD = 3,378) in the HRS.

### 4.4. Analysis Methods

As described above in Subsections 3.2 and 3.3, interviewer-days are cross-classified within interviewers and PSUs in both the NSFG and HRS. Thus, we use cross-classified multilevel regression models, with interviewer-days nested within interviewers and PSUs, to examine the association between field effort and interviewer travel. In all models, we test whether a random effects model is needed using a likelihood ratio test that is a mixture of chi-squared distributions (West et al. 2015, 107).

#### 4.4.1. Models for Travel

For the first research question, we examine predictors of the number of miles traveled and the number of segments visited.

*Model for Miles*

First, in order to predict the number of miles traveled in the NSFG we estimate a hierarchical linear regression model with a normal distribution and identity link function:

$$
\begin{aligned}
1\mathrm{pc}miles_{i(jk)} = \theta_0 &+ \beta_1 segments_{i(jk)} + \beta_2 SmallMSA_k + \beta_3 NonMSA_k + \beta_4 CensusHTC_k \\
&+ \beta_5 AreaSqMiles_k + \beta_6 AnyExp_j + \beta_7 Week_{i(jk)} + \beta_8 Phase1_{i(jk)} \\
&+ \beta_9 Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k} + e_{ijk}
\end{aligned}
$$

where $i$ represents interviewer-days, $j$ represents interviewers, $k$ represents PSUs, *miles* is the total number of miles traveled per interviewer-day, *segments* is the total number of segments visited per interviewer-day, *SmallMSA* and *NonMSA* are dichotomous indicator variables for the urbanicity of the PSU, *CensusHTC* is a continuous measure indicating the Census Hard-to-Count score (Bruce and Robinson 2006), *AreaSqMiles* is the centered number of square miles in the PSU, *AnyExp* is an interviewer-level indicator for whether the interviewer has any prior interviewing experience, *Week* is a continuous measure of the week of the field period, *Phase1* is an indicator variable for the interviewer-day being early in the field period, $b_{00j}$ is a random effect for interviewers with a normal distribution and mean of zero, $c_{00k}$ is a random effect for PSUs with a normal distribution and mean of zero, and $e_{ijk}$ is a residual term (Raudenbush and Bryk 2002). The model in the HRS is identical except that the two urbanicity variables are replaced with one indicator variable for whether the segment is located in a self-representing PSU or not. SAS 9.4 PROC MIXED is used to estimate these hierarchical linear models.

*Model for Segments*

We estimate a hierarchical Poisson model with a log link to predict the number of segments visited:

$$\begin{aligned}\log(segments_{ij}) = \theta_0 &+ \beta_1 miles_{i(jk)} + \beta_2 SmallMSA_k + \beta_3 NonMSA_k \\ &+ \beta_4 CensusHTC_k + \beta_5 AreaSqMiles_k + \beta_6 AnyExp_j + \beta_7 Week_{i(jk)} \\ &+ \beta_8 Phase1_{i(jk)} + \beta_9 Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k}\end{aligned}$$

where the predictors are as defined above. The number of miles driven per day is added as a predictor to this model. All Poisson models are estimated using SAS 9.4 PROC GLIMMIX.

4.4.2. Models for Field Outcomes

For the second research question, we examine whether there is a relationship between travel and field outcomes.

*Model for Attempts*

When examining the total number of contact attempts for the screener or the main interview, we estimate a hierarchical negative binomial regression model with a log link function:

$$\begin{aligned}\log(attempts_{ij}) = \theta_0 &+ \beta_1 segments_{i(jk)} + \beta_2 miles_{i(jk)} + \beta_3 SmallMSA_k + \beta_4 NonMSA_k \\ &+ \beta_5 CensusHTC_k + \beta_6 AreaSqMiles_k + \beta_7 AnyExp_j + \beta_8 Week_{i(jk)} \\ &+ \beta_9 Phase1_{i(jk)} + \beta_{10} Phase1_{i(jk)} Week_{i(jk)} + b_{00j} + c_{00k}\end{aligned}$$

We initially estimated hierarchical Poisson models. These models had very poor model fit with evidence of overdispersion (Generalized Chi-Square/DF $>$ 2 for all models; Stroup 2011). Thus, hierarchical negative binomial models were estimated. The negative binomial models significantly improved model fit over the Poisson models for the total number of contact attempts (Generalized Chi-Square $\sim\, =1$ in both surveys; statistically significant scale parameters). All negative binomial models are estimated using SAS 9.4 PROC GLIMMIX.

*Models for Contact and Interview Rates*

For the contact, screener and main interview rates, we estimate a hierarchical Poisson regression model predicting the number of contact attempts with successful contacts, screener and main interviews with the number of contact attempts, contact attempts to obtain a screener, and contact attempts to obtain a main interview as the offset variables, respectively. All Poisson models are estimated using SAS 9.4 PROC GLIMMIX. For example, for the contact rate, we estimate:

$$\log(contacts/visits_{ij}) = \theta_0 + \beta_1 segments_{i(jk)} + \beta_2 miles_{i(jk)} + \beta_3 SmallMSA_k + \beta_4 NonMSA_k + \beta_5 CensusHTC_k + \beta_6 AreaSqMiles_k + \beta_7 AnyExp_j + \beta_8 Week_{i(jk)} + \beta_9 Phase1_{i(jk)} + \beta_{10} Phase1_{i(jk)}\,Week_{i(jk)} + b_{00j} + c_{00k}$$

For these models, we use the same predictors as the previous models, and include both segments visited and miles travelled per day as predictors in these models.

## 5. Findings

### 5.1. What are Predictors of Interviewer Travel Behavior in Two Large National US Field Surveys?

We now turn to our first research question – what predicts interviewer travel behavior in these two surveys? We start by estimating a base model predicting mileage with no predictors. In the NSFG, there is a significant variance component related to interviewers ($var(b_{00j}) = 7157.1$) and to the PSUs ($\mathrm{var}(c_{00k}) = 5154.7$; likelihood ratio chi-square = 1691.6, $p < .0001$; see the top panel of Table 2), with an interviewer intraclass correlation coefficient of 51.3% and a PSU intraclass correlation coefficient of 36.9%. There is also a significant variance component related to interviewers in the HRS ($var(b_{00j}) = 2397.1$) and to the PSUs ($\mathrm{var}(c_{00k}) = 61.2$; likelihood ratio chi-square = 11963.5, $p < .0001$; see the top panel Table 2), with an interviewer intraclass correlation coefficient of 71.8% and a PSU intraclass correlation coefficient of 1.8%. This means that over 50% of the variance in mileage traveled is due to interviewers overall in both surveys, but that the variance in mileage due to PSUs differs dramatically across the two surveys.

We note that interviewers vary in the characteristics of their assigned PSUs, their skill sets, as well as the proximity of their home to sampled segments. These factors may explain some or all of the variation between interviewers. To address this, we now look at predictors of the number of miles traveled each interviewer-day.

As shown in the top panel of Table 2, the total distance traveled is significantly associated with the number of segments visited in the NSFG, but not the HRS. In the NSFG, interviewers travel almost eight miles more for each additional segment visited. Counter to our expectations, there is no systematic linear association between urbanicity, the Census Hard-to-Count score, the size of the area segment, interviewer experience, or time in the field period and mileage in the NSFG.

The associations in the HRS are somewhat different than in the NSFG. In the HRS, there is no systematic association between the number of segments visited and mileage. As in the NSFG, and again counter to our expectations, there is no association between the Census Hard-to-Count score, the area of the PSU, urbanicity, interviewer experience, or the time in the field period with miles traveled in the HRS.

*Table 2. Cross-Classified Random Effects Linear Regression Coefficients and Standard Errors Predicting Number of Miles Travelled Per Interviewer-Day and Cross-Classified Random Effects Poisson Regression Coefficients and Standard Errors Predicting Number of Segments Visited Per Interviewer-Day, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).*

| | NSFG | | | HRS | | |
|---|---|---|---|---|---|---|
| | Base Model | Full Model | | Base Model | Full Model | |
| | | Coefficient | Standard Error | | Coefficient | Standard Error |
| Miles – Linear Regression | | | | | | |
| Intercept | 92.49**** | 55.96 | 88.75 | 53.19**** | 61.48 | 37.18 |
| Number of segments | | 7.98**** | 0.95 | | 0.26 | 0.24 |
| Small MSA | | 6.69 | 35.01 | | – | |
| Non-MSA | | 44.34 | 44.91 | | – | |
| Non-self-representing PSU | | – | | | −2.05 | 1.93 |
| Census Hard-to-Count Score | | −1.11 | 1.37 | | 0.08 | 0.07 |
| Area of PSU (Square Miles) | | 0.009 | 0.01 | | −0.0005 | 0.0003 |
| Any experience | | 3.75 | 45.09 | | 5.91 | 7.75 |
| Week in the field period | | 4.36 | 5.31 | | −0.47 | 0.73 |
| Phase 1 | | 40.51 | 60.68 | | −2.05 | 36.44 |
| Week in the field period*Phase 1 | | −4.00 | 5.34 | | 0.06 | 0.73 |
| Variance Components | | | | | | |
| Interviewer ($var(b_{00j})$) | 7157.1**** | 7450.8**** | | 2397.1**** | 2378.5**** | |
| Area ($var(c_{00k})$) | 5154.7*** | 5858.6** | | 61.2**** | 70.7**** | |
| Residual ($var(e_{ij})$) | 1646.4**** | 1559.5**** | | 881.9**** | 866.9**** | |
| Chi-square test for variance components | 1691.6**** | 1509.0**** | | 11963.5**** | 11649.9**** | |
| AIC | 16899.9 | 16777.3 | | 120822.6 | 120144.6 | |
| −2 Log-Likelihood | 16893.3 | 16771.3 | | 120816.6 | 120138.6 | |

*Table 2. Continued.*

| | NSFG | | | HRS | | |
|---|---|---|---|---|---|---|
| | Base Model | Full Model | | Base Model | Full Model | |
| | | Coefficient | Standard Error | | Coefficient | Standard Error |
| Segments – Poisson regression | | | | | | |
| Intercept | 0.62**** | 3.80** | 1.10 | 0.61**** | 3.88*** | 0.99 |
| Number of miles | | 0.002**** | 0.0003 | | 0.0003 | 0.0002 |
| Small MSA | | −0.06 | 0.09 | | – | |
| Non-MSA | | 0.05 | 0.11 | | – | |
| Non-self-representing PSU | | – | | | 0.10* | 0.04 |
| Census Hard-to-Count Score | | 0.005 | 0.003 | | −0.002 | 0.001 |
| Area of PSU (Square Miles) | | −0.00007* | 0.00003 | | −0.00001 | 0.000006 |
| Any experience | | −0.06 | 0.12 | | 0.13** | 0.04 |
| Week in the field period | | −0.30** | 0.005 | | −0.07*** | 0.02 |
| Phase 1 | | −3.79*** | 1.09 | | −3.47*** | 0.99 |
| Week in the field period*Phase 1 | | 0.36*** | 0.09 | | 0.07*** | 0.02 |
| Interviewer (*var*($b_{00j}$)) | 0.022 | 0.019 | | 0.049**** | 0.046**** | |
| Area (*var*($c_{00k}$)) | 0.031* | 0.021 | | 0.029**** | 0.019**** | |
| Chi-square test for variance components | 72.5**** | 39.4**** | | 1216.7**** | 954.6**** | |
| −2 Log Pseudo-Likelihood | 3410.0 | 2986.5 | | 23995.3 | 22952.4 | |
| Generalized Chi-Square | 1133.4 | 857.6 | | 7817.6 | 7461.0 | |
| Generalized Chi-Square/DF | 0.64 | 0.53 | | 0.61 | 0.60 | |

*Note:* Models account for clustering of interviewer-days within interviewer and within area, $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$. AIC = Akaike Information Criterion; DF = Degrees of freedom; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.

Next, we look at predictors of the number of segments visited on each interviewer day (bottom panel of Table 2). Here, the predictors are surprisingly different from those of mileage. Mirroring the model predicting the number of miles driven, in the NSFG, the number of miles driven is positively associated with the number of segments visited, but in the HRS this coefficient is not significantly different from zero. The urbanicity indicators are not associated with the number of segments visited per day in the NSFG, counter to our expectations, but interviewers who work in non-self-representing PSUs visit more segments per day, on average, than interviewers who work in other PSUs in the HRS, as expected. In the NSFG, experienced and inexperienced interviewers visit the same number of segments per day, on average, but in the HRS, experienced interviewers visit more segments on average than inexperienced interviewers each day. This is likely due to constraints in the number of segments assigned to interviewers in the NSFG (only three), whereas the HRS interviewers may have larger numbers of assigned segments. In both surveys, fewer segments are visited earlier in the field period than later in the field period, indicating more cross-segment travel late in the field period, and the rate of change across the weeks for the first and second phases also changes.

### 5.2. *Is Interviewer Travel Associated with Field Outcomes?*

We now look at the relationship between these two travel measures and field behaviors. We start with the total number of contact attempts. There is significant variation across interviewers and PSUs in both studies in the number of screener interview contact attempts (base models in Table 3, NSFG: $\mathrm{var}(b_{00j}) = 0.10$, $\mathrm{var}(c_{00k}) = 0.06$, $p < .0001$; HRS: $\mathrm{var}(b_{00j}) = 0.32$, $\mathrm{var}(c_{00k}) = 0.18$, $p < .0001$) and main interview contact attempts (NSFG: $\mathrm{var}(b_{00j}) = 0.08$, $\mathrm{var}(c_{00k}) = 0.13$ $p < .0001$; HRS: $\mathrm{var}(b_{00j}) = 0.32$, $\mathrm{var}(c_{00k}) = 0.18$) per day. As shown in Table 3 in both surveys, as expected, as the number of segments visited increases, the number of screener and main contact attempts made during an interviewer-day also increases (NSFG screener $b = 0.25$, $p < .0001$; NSFG main $b = 0.20$, $p < .0001$; HRS screener $b = 0.26$, $p < .0001$; HRS main $b = 0.27$, $p < .0001$). Strikingly and surprisingly, there is no relationship ($p > 0.05$) between the number of miles traveled and the total number of main attempts in either survey, and there is a weak association ($b = 0.001$, $p < .05$) between the number of miles and screener attempts in the NSFG, but not the HRS. There is no association between urbanicity and contact attempts across the two surveys, and no association between an interviewer's prior experience, the Census Hard-to-Count score, and the area of the PSU, and the number of contact attempts. In both surveys, the number of screener attempts decreases as the field period progresses, although the decrease is stronger in Phase 2 of the survey. The number of main interview attempts does not change as the field period progresses in the HRS ($p > 0.05$), but increases late in the field period in the NSFG ($p < .01$). The NSFG has a short field period, and although the survey managers emphasize completing screening interviews early in the field period, there is a push toward completing main interviews at the end of the field period. For the HRS, the interview is much longer. In many cases, two persons within the household are interviewed. These conditions made it more difficult to schedule and complete these interviews.

*Table 3. Cross-Classified Random Effects Negative Binomial Regression Coefficients and Standard Errors Predicting Number of Contact attempts Per Interviewer-Day, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).*

| | NSFG | | | | | | HRS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Screener | | | Main | | | Screener | | | Main | | |
| | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | |
| | | Coefficient | Standard Error | | Coefficient | Standard Error | | Coefficient | Standard Error | | Coefficient | Standard Error |
| Intercept | 2.30**** | 8.69**** | 1.49 | 1.37**** | 2.99* | 1.26 | 2.15**** | 59.17**** | 7.34 | 0.069 | −1.08 | 1.31 |
| Distance in miles | | 0.001* | 0.0004 | | 0.0005 | 0.0004 | | 0.0002 | 0.0003 | | −0.0007 | 0.0004 |
| Number of segments visited | | 0.25**** | 0.02 | | 0.20**** | 0.02 | | 0.26**** | 0.007 | | 0.27**** | 0.009 |
| Small MSA | | −0.26 | 0.16 | | 0.20 | 0.16 | | | | | | |
| Non-MSA | | −0.41 | 0.21 | | 0.17 | 0.21 | | | | | | |
| Non-self-representing PSU | | – | | | | | | −0.11 | 0.06 | | −0.01 | 0.07 |
| Census Hard-to-Count Score | | 0.01 | 0.006 | | 0.0006 | 0.006 | | 0.005 | 0.003 | | −0.003 | 0.002 |
| Area of PSU (Square Miles) | | −0.00003 | 0.00006 | | 0.00008 | 0.00006 | | 0.00001 | 0.00001 | | 0.000003 | 0.00001 |
| Any experience | | 0.03 | 0.25 | | −0.10 | 0.21 | | 0.07 | 0.09 | | 0.09 | 0.11 |
| Week in the field period | | −0.75**** | 0.13 | | −0.23* | 0.11 | | −1.25**** | 0.15 | | −0.01 | 0.02 |
| Phase 1 | | −6.24**** | 1.46 | | −2.39 | 1.22 | | −57.43**** | 7.34 | | −2.18 | 1.30 |
| Week in the field period*Phase 1 | | 0.60**** | 0.13 | | 0.30** | 0.11 | | 1.24**** | 0.15 | | 0.04 | 0.03 |

*Table 3. Continued.*

| | NSFG | | | | | | HRS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Screener | | | Main | | | Screener | | | Main | | |
| | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | |
| | | Coefficient | Standard Error | | Coefficient | Standard Error | | Coefficient | Standard Error | | Coefficient | Standard Error |
| Dispersion | 0.86**** | 0.48**** | | 0.49**** | 0.22**** | | 0.70**** | 0. 642**** | | 0.82**** | 0.61**** | |
| Interviewer ($var(b_{00j})$) | 0.10** | 0.17**** | | 0.08** | 0.10*** | | 0.32**** | 0.24**** | | 0.52**** | 0.37**** | |
| Area ($var(c_{00k})$) | 0.06 | 0.08* | | 0.13** | 0.10** | | 0.18**** | 0.15**** | | 0.16**** | 0.11**** | |
| *Chi-square test for variance components* | 148.0**** | 197.8**** | | 218.5**** | 237.0**** | | 2221.7**** | 1787.2**** | | 1471.1**** | 1137.5**** | |
| −2 Log-Likelihood | 5098.1 | 4049.0 | | 4690.0 | 3701.6 | | 34836.3 | 152984.9 | | 44150.6 | 41714.5 | |
| Generalized Chi-square | 1783.8 | 1659.7 | | 1761.7 | 1614.2 | | 12985.0 | 131668.3 | | 12770.5 | 12678.1 | |
| Generalized Chi-Square/DF | 1.00 | 1.03 | | 0.99 | 1.00 | | 1.01 | 10.63 | | 0.99 | 1.02 | |

*Note:* Models account for clustering of interviewer-days within interviewer, $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$. DF = Degrees of Freedom; MSA = Metropolitan Statistical Area; PSU = Primary Sampling Unit.

We now examine the relationship between travel behavior and contact rates (Table 4). As with the other field outcomes, there is significant variability across interviewers and PSUs in contact rates (NSFG: var($b_{00j}$) = 0.09, var($c_{00k}$) = 0.04, $p < .0001$; HRS: var($b_{00j}$) = 0.08, var($c_{00k}$) = 0.02, $p < .0001$). Consistent with our expectation that interviewer travel and contact rates will be negatively correlated, Table 4 shows that there is a modest, but noticeable decline in contact rates in both surveys as an interviewers' travel increases as measured by the number of segments (NSFG: $b = -0.03$, $p < .001$; HRS: $b = -0.03$, $p < .0001$). Unexpectedly, we see no relationship at all between the distance traveled in miles and contact rates ($p > 0.05$). Non-urban PSUs as represented by the non-MSAs in the NSFG ($b = 0.36$, $p < .01$) and by non-self-representing PSUs in the HRS ($n = 0.07$, $p < .0001$) have higher contact rates, on average, than urban PSUs in both surveys. We see no difference in contact rates for interviewers with some prior experience compared to those without prior experience in either survey. Contact rates do not systematically change over the field period. There is no association between the Census Hard-to-Count score and contact rates ($p > 0.05$). Larger PSUs have higher contact rates in the NSFG ($b = 0.0001$, $p < .01$), but not in the HRS ($p > 0.05$).

The next outcomes, also presented in Table 4, are screener interview rates and main interview rates. There is significant variability across interviewers in screening and main interview rates (NSFG: screening var($b_{00j}$) = 0.11, var($c_{00k}$) = 0.08, $p < .0001$, main var($b_{00j}$) = 0.09, var($c_{00k}$) = 0.05, $p < .0001$; HRS: screening var($b_{00j}$) = 0.16, var($c_{00k}$) = 0.05, $p < .0001$, main var($b_{00j}$) = 0.17, var($c_{00k}$) = 0.09, $p < .0001$). We anticipate the same association as described with contact rates – a negative association between travel (as measured by miles and number of segments visited) and screener / main interview rates, with the same rationale. As with contact rates and consistent with our expectations, there is a statistically significant negative association between screener rates and the total number of segments visited on an interviewer-day and also between main interview rates and the total number segments visited on an interviewer-day in both surveys (coefficients range from $-0.05$ to $-0.27$, $p < .001$). Neither survey displays a significant relationship ($p < .05$) between the total distance traveled in miles and any of these field outcomes. Thus, the measure of travel that predicts these important field outcomes (and error indicators) in both surveys is how many different segments are visited, not the number of miles that an interviewer drives. Non-urban PSUs tend to have higher screener and main interview rates (NSFG: $b = 0.55$, $p < .01$ screener; $b = 0.30$, $p = 0.06$ main; HRS: $b = 0.12$, $p < .0001$ screener; $b = 0.41$, $p < .0001$ main) in both surveys. There is no association between interviewer-day level screener interview and main interview rates and interviewer experience. For the HRS, interviewer-day level screener interview rates changes as the field period progresses. In both surveys, main interview rates change as the field period progresses. There is no association between the Census Hard-to-Count score and screener or main completion rates. Screener completion rates are associated with larger PSUs in the NSFG ($b = 0.005$, $p < .01$), but no association with main interview rates, and no association in the HRS ($p > 0.05$).

## 6. Summary and Discussion

Although costs of interviewer travel have been examined with respect to sample designs (e.g., Kalsbeek et al. 1994; Bienias et al. 1990) predictors of interviewer travel behavior

*Table 4. Cross-Classified Random Effects Poisson Regression Coefficients and Standard Errors Predicting Interviewer-Day Level Contact, Screener Interview and Main Interview Rates, for the National Survey of Family Growth (NSFG) and Health and Retirement Study (HRS).*

| | NSFG | | | | | | | | | HRS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Contact | | | Screener | | | Main | | | Contact | | | Screener | | | Main | | |
| | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | | Base Model | Full Model | |
| | | Coef. | SE | | Coef. | SE | | Coef. | SE | | Coef. | SE | | Coef. | SE | | Coef. | SE |
| Intercept | −0.10**** | −2.52 | 1.26 | −1.61**** | −2.29 | 2.80 | −1.59**** | −6.04* | 2.48 | −1.04**** | −1.45**** | 1.20 | −1.51**** | −1.10**** | 0.10 | −1.90**** | −1.10**** | 0.17 |
| Distance in miles | | −0.0003 | 0.0002 | | 0.000005 | 0.0004 | | 0.0002 | 0.0004 | | −0.0002 | 0.0001 | | −0.0003 | 0.0002 | | 0.0005 | 0.0005 |
| Number of segments | | −0.03** | 0.01 | | −0.05** | 0.02 | | −0.18**** | 0.02 | | −0.03**** | 0.004 | | −0.06**** | 0.005 | | −0.27**** | 0.02 |
| Small MSA | | 0.03 | 0.10 | | 0.05 | 0.14 | | 0.05 | 0.13 | | – | | | – | | | | |
| Non-MSA | | 0.36** | 0.13 | | 0.55** | 0.17 | | 0.30 | 0.16 | | – | | | – | | | | |
| Non-self-representing PSU | | | | | – | | | – | | | 0.07**** | 0.02 | | 0.12**** | 0.05 | | 0.41**** | 0.10 |
| Census Hard-to-Count Score | | −0.006 | 0.004 | | −0.010 | 0.005 | | −0.009 | 0.005 | | −0.002 | 0.001 | | −0.002 | 0.002 | | 0.0007 | 0.003 |
| Area of PSU (Square Miles) | | 0.0001** | 0.00004 | | 0.0001** | 0.00005 | | 0.00003 | 0.00004 | | −0.000003 | 0.000006 | | 0.000005 | 0.000009 | | 0.000004 | 0.00001 |
| Any experience | | −0.16 | 0.17 | | −0.09 | 0.20 | | −0.21 | 0.18 | | 0.03 | 0.05 | | −0.04 | 0.07 | | −0.11 | 0.09 |
| Week in the field period | | 0.15 | 0.11 | | 0.04 | 0.25 | | 0.42* | 0.22 | | 0.01 | 0.02 | | −0.008**** | 0.0006 | | −0.02**** | 0.002 |
| Phase 1 | | 2.03 | 1.24 | | 1.35 | 2.79 | | 5.55* | 2.47 | | 0.70 | 1.19 | | 16.14**** | 4.09 | | 4.76 | 3.61 |
| Week in the field period*Phase 1 | | −0.16 | 0.11 | | −0.11 | 0.25 | | −0.47* | 0.22 | | −0.02 | 0.02 | | 0.32**** | 0.08 | | −0.11 | 0.07 |
| *Interviewer var($b_{00j}$)* | 0.09*** | 0.08**** | | 0.11*** | 0.10**** | | 0.09** | 0.05* | | 0.08**** | 0.08**** | | 0.16**** | 0.14**** | | 0.17**** | 0.12**** | |
| *Area var($c_{00k}$)* | 0.04** | 0.03** | | 0.08* | 0.05* | | 0.05 | 0.04 | | 0.02**** | 0.02**** | | 0.05**** | 0.05**** | | 0.09**** | 0.09**** | |
| *Chi-square test for variance components* | 702.7**** | 439.4**** | | 409.2**** | 199.8**** | | 104.6**** | 50.9**** | | 3391.3**** | 3127.6**** | | 3248.0**** | 2664.3**** | | 358.6**** | 261.6**** | |
| −2 Log Pseudo-Likelihood | 3514.45 | 2974.51 | | 3929.72 | 3689.6 | | 5657.69 | 5029.50 | | 27743.18 | 25938.50 | | 32325.02 | 30392.78 | | 31209.06 | 29742.38 | |
| Generalized Chi-square | 2220.44 | 1915.30 | | 1801.54 | 1624.2 | | 2002.76 | 1713.95 | | 14966.36 | 14072.51 | | 14546.67 | 13622.59 | | 8632.88 | 7840.19 | |
| Generalized Chi-Square/DF | 1.25 | 1.19 | | 1.23 | 1.18 | | 1.25 | 1.16 | | 1.16 | 1.14 | | 1.24 | 1.20 | | 1.21 | 1.13 | |

*Note:* Models account for clustering of interviewer-days within interviewer and area, *$p < .05$, **$p < .01$, ***$p < .001$, ****$p < .0001$; Total number of contact attempts for a contact, screener or main interview used as offset in each model. DF = Degrees of Freedom; Coef = Coefficient; SE = Standard Error; PSU = Primary Sampling Unit; MSA = Metropolitan Statistical Area.

and the association between travel decisions and field outcomes to date has not received any attention. In this article, we conducted an initial examination of variables that may be associated with interviewer travel, treating indicators of interviewer travel behavior as outcomes in some models and predictors in others. In the models, geographic characteristics such as the size in square miles of the PSU and urbanicity were not associated with travel outcomes such as the numbers of miles traveled or segments visited. In models predicting the number of segments visited, characteristics of the data collection, such as design phases or the week of data collection, were associated with these travel outcomes. In these models, interviewer variance was a significant component of the variance. In terms of field outcomes, the geographic size of the PSU did not play a significant role except for screener and contact rates in the NSFG. The week of the field period and the design phase were associated with the number of attempts made in a day.

In terms of the relationship of interviewer travel to the six different field outcomes and error indicators, we see a clear, consistent pattern for the two surveys, summarized in Table 5. The associations are clear – interviewers who visit more segments on a given day also have more contact attempts and have lower contact and response rates. This effect holds even accounting for the number of miles that the interviewers travel. In none of the analyses is the raw number of miles traveled by interviewers associated with the number of contact attempts or contact or response rates.

The replication of the findings about the association (or lack thereof) between overall distance and number of segments visited and field outcomes despite the design differences between the two surveys is encouraging. The field period in the HRS and NSFG is very different. In the NSFG, there is a limited (twelve week) field period. Interviewers are encouraged to visit every sampled housing unit as quickly as possible and visit as many segments as they can every day. In this way, the NSFG encourages interviewers to maximize their travel. In contrast, the HRS has an extended (twelve month) field period, and interviewers are encouraged to minimize their travel as the field period progresses. These differences in field period and survey management yield differences in the relationship between travel behaviors and contact attempts over the field period across the two surveys; yet there are few differences in the association between travel and field outcomes between these two surveys. The number of segments an interviewer visits has a positive association with contact attempts, but negative association with response rates. These results suggest that this finding may generalize across survey settings.

*Table 5. Summary of association between travel and field outcomes.*

| | Is there an association between travel and field outcome? | |
|---|---|---|
| Field Outcome | Miles | Segments |
| Screener contact attempts | Weak (+ NSFG only) | Yes (+) |
| Main contact attempts | No | Yes (+) |
| Contact rates | No | Yes (−) |
| Screener rates | No | Yes (−) |
| Main interview rates | No | Yes (−) |

*Note:* The sign in parentheses in the "segments" column indicates the direction of the association. NSFG = National Survey of Family Growth.

The association between interviewer travel behavior and contact and cooperation rates is important for survey practice for three reasons. First, most survey organizations closely monitor response rates, potentially indicating differential nonresponse bias across interviewers (West and Olson 2010). Although Groves and Peytcheva (2008) found that the response rate may not be a good indicator for the risk of bias, it still is a baseline quality measure used to assess interviewers by many survey organizations. Second, the ability to control variability between interviewers in the response rate is an important prerequisite in controlling differential response rates across subgroups, which is a strategy for minimizing the risk of nonresponse bias (Montaquila et al. 2008; Groves 2006; Schouten et al. 2016). Finally, we do not know whether different types of travel decisions are associated with contact or cooperation rates. Thus, establishing whether such an association exists is a critical first step for developing interviewer training related to travel and for understanding variability in interviewer response rates. Our findings imply that survey organizations should carefully monitor interviewer travel behavior – and in particular, between segment travel – as a way of reducing between-interviewer variability in response rates, and thus minimize the risk of nonresponse error variance due to the interviewer.

These findings also have important implications for costs and practice by survey field managers. Although interviewer mileage should be monitored for a simple cost calculation (travel costs $=$ number of miles $*$ reimbursement rate per mile), the number of segments visited each day should also be monitored as an additional indicator of survey error and cost. Furthermore, we recommend that the number of segments visited on a given day be used to initiate conversations with interviewers about their travel. In particular, field managers should investigate why interviewers are traveling to additional segments as the travel is an indication that calling is less productive on those trips. This may be due to the time of day and day of the week of the trip, the approach of the interviewer, or other factors which may need further investigation. Thus, the number of segments could be a useful way to monitor and provide feedback to interviewers on their behavior and obtain new insights into the reasons for nonparticipation.

This article is not without limitations. First, establishing causality from existing administrative travel data is difficult. Most travel data is reported at an aggregate interviewer-day level in timesheets; that is, travel data is not associated with individual cases or contact attempts but instead with each day that an interviewer works. Thus, we can examine associations in this analysis, but cannot determine whether different travel behaviors cause field outcomes, or field outcomes cause different travel behavior. The observed associations could also be the result of particular allocations of sample to interviewers. For example, if more experienced interviewers are allocated more difficult samples, this could lead to experience being less predictive in the estimated models while it might actually be the case that experienced interviewer produce higher response rates than less experienced interviewers when they are allocated equivalent samples. This type of allocation based on difficulty was not the usual practice in either of the surveys used here, but might be used by other surveys. Second, these two surveys are large, national surveys with screening to find particular target populations. Each of these surveys also oversamples area segments with higher proportions of black and Hispanic persons. We do not know how these results would replicate in surveys without this screening step and oversampling. Third, the segments were not randomly assigned to interviewers, and as such, all of our findings are from an observational study. We have attempted to account for these potential differences in interviewer assignment PSUs using cross-classified random effects models and including additional area-level predictors (e.g., see Stokes and Jones 1989), but we have not explained away the PSU effect. Additionally, neither of these surveys pay interviewers per completed interview, nor require interviewers to work multiple studies at one time. Instead, NSFG and HRS pays interviewers by the hour, and both hire interviewers only for one study at a time. It is not clear whether surveys that use a different pay structure would see similar patterns.

Future evaluations of the quality of travel data are needed. Although we have initial indications that interviewer-reported mileage and mileage calculated from geocoded call records differ, we do not know which source is more accurate. Given that all of the analyses replicated with both timesheet and geocoded mileage information, we believe that our lack of association with mileage is robust to measurement errors in the timesheet data. GPS devices allow the collection of data regarding the location of interviewers. These data include latitude, longitude, speed, direction, altitude, and time and date. Collection of real time travel data through the use of GPS devices carried by interviewers would help us understand the quality of both of these sources of data (Olson and Wagner 2015; Wagner et al. 2013). Additionally, collection of real time travel data would permit the examination of the relationship between travel and field outcomes at the address level, rather than the interviewer-day level. It would also allow us to evaluate the amount of time spent to travel a certain number of miles. Interviewers may take a longer route at which they can travel with faster speed such as on a freeway, thus taking less time from their interviewing day than a shorter route at slower speeds.

Future research will incorporate information about both the distance traveled and the number of segments visited per visit into explicit cost models. Although interviewer travel is often mentioned as a constraint on the number of clusters to select and the size of the clusters, we are unaware of cost-error models that incorporate empirical data for these measures. The data presented here could provide useful inputs for future cost models related to interviewer travel in face-to-face surveys.

## 7. References

Biemer, P.P., P. Chen, and K. Wang. 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. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01058.x.

Bienias, J.L., E.M. Sweet, and C.H. Alexander. 1990. A Model for Simulating Interviewer Travel Costs for Different Cluster Sizes. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–27. Anaheim, CA. Available at: https://ww2.amstat.org/sections/srms/Proceedings/papers/1990_004.pdf (accessed October 14, 2017).

Blom, A.G. 2012. "Explaining Cross-Country Differences in Survey Contact Rates: Application of Decomposition Methods." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1): 217–242. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.01006.x.

Blom, A.G., E.D. de Leeuw, and J.J. Hox. 2011. "Interviewer Effects on Nonresponse in the European Social Survey." *Journal of Official Statistics* 27(2): 359–377.

Bruce, A. and J.G. Robinson. 2006. "Tract-Level Planning Database with Census 2000 Data." US Department of Commerce, US Census Bureau: Washington, DC, 226. Available at: https://www.census.gov/2010census/partners/pdf/TractLevelCensus-2000Apr_2_09.pdf (accessed October 14, 2017).

Campanelli, P., P. Sturgis, and S. Purdon. 1997. *Can You Hear Me Knocking?: Investigation into the Impact of Interviewers on Survey Response Rates*, National Centre for Social Research, London.

Chen, B.-C. 2012. "Simulating NHIS Field Operations." *Proceedings of the 2012 Research Conference*, Federal Committee on Statistical Methodology. Office of Management and Budget. Washington, DC. Available at: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/Chen_2012FCSM_II-A.pdf (accessed October 14, 2017).

Cochran, W.G. 1977. *Sampling Techniques* (Third ed.). New York: John Wiley and Sons.

Durrant, G.B. and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(2): 361–381. Doi: http://dx.doi.org/10.1111/j.1467-985X.2008.00565.x.

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5): 646–675. Doi: http://dx.doi.org/10.1093/poq/nfl0.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72(2): 167–189.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. 1953. *Sample Survey Methods and Theory*. New York: Wiley.

Health and Retirement Study. 2008. *Sample Evolution: 1992–1998*. Available at: http://hrsonline.isr.umich.edu/sitedocs/surveydesign.pdf (accessed February 6, 2014).

Heeringa, S.G. and J.H. Connor. 1995. *Technical Description of the Health and Retirement Survey Sample Design*. Available at: http://hrsonline.isr.umich.edu/sitedocs/userg/HRSSAMP.pdf (accessed February 6, 2014).

Judkins, D., J. Waksberg, and D. Northrup. 1990. Cost Functions for NHIS and Implications for Survey Design. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 34–43. Anaheim, CA.

Kalsbeek, W.D., O.M. Mendoza, and D.V. Budescu. 1983. "Cost Models for Optimum Allocation in Multi-Stage Sampling." *Survey Methodology* 9(2): 154–177.

Kalsbeek, W.D., S.L. Botman, J.T. Massey, and P.-W. Liu. 1994. "Cost-Efficiency and the Number of Allowable Call Attempts in the National Health Interview Survey." *Journal of Official Statistics* 10(2): 133–152.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lepkowski, J.M., W.D. Mosher, K.E. Davis, R.M. Groves, and J. Van Hoewyk. 2010. *The 2006–2010 National Survey of Family Growth: Sample design and analysis of a continuous survey*. Washington, DC: National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/data/series/sr_02/sr02_150.pdf (accessed October 14, 2017).

Mayer, C.S. 1968. "A Computer System for Controlling Interviewer Costs." *Journal of Marketing Research* 5: 312–318.

Montaquila, J.M., J.M. Brick, M.C. Hagedorn, C. Kennedy, and S. Keeter. 2008. "Aspects of Nonrepsonse Bias in RDD Telephone Surveys." In *Advances in Telephone Survey*

*Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster, 561–586. Hoboken, NJ: John Wiley & Sons.

Morton-Williams, J. 1993. *Interviewer Approaches*. Hants, England: Aldershot.

Olson, K. and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71(2): 273–286.

Olson, K. and J. Wagner. 2015. "A Feasibility Test of Using Smartphones to Collect GPS Information in Face-to-Face Surveys." *Survey Research Methods* 9(1): 1–13. Doi: http://dx.doi.org/10.18148/srm/2015.v9i1.6036.

O'Muircheartaigh, C. and P. Campanelli. 1999. "A Multilevel Exploration of the Role of Interviewers in Survey Non-Response." *Journal of the Royal Statistical Society, Series A* 162: 437–446. Doi: http://dx.doi.org/10.1111/1467-985X.00147.

Peachman, J. 1992. *Design and Methodology of 1991/92 Household Interview Survey*. Paper presented at the Australasian Transport Research Forum 17th Conference, 149–161. Canberra, Australia. Available at: http://atrf.info/papers/1992/1992_Peachman.pdf (accessed October 14, 2017).

Pickery, J. and G. Loosveldt. 2002. "A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse." *Quality and Quantity* 36: 427–437. Doi: https://doi.org/10.1023/A:1020905911108.

Purdon, S., P. Campanelli, and P. Sturgis. 1999. "Interviewers' Calling Strategies on Face-to-Face Interview Surveys." *Journal of Official Statistics* 15(2): 199–219.

Raudenbush, S.W. and A.S. Bryk. 2002. *Hierarchical linear models : applications and data analysis methods*. Thousand Oaks, Sage Publications.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society: Series A* (Statistics in Society) 179(3): 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. West Sussex, UK: Wiley.

Stokes, S.L. and P.A. Jones. 1989. Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey. *1989 Proceedings of the Survey Research Methods Section*. American Statistical Association. Alexandria, VA, 696–698.

Stroup, W. 2011. Living with Generalized Linear Models. Paper 349-2011, SAS Global Forum 2011. Available at: http://support.sas.com/resources/papers/proceedings11/349-2011.pdf (accessed October 14, 2017).

Sudman, S. 1965-66. "Time Allocation in Survey Interviewing and in Other Field Occupations." *Public Opinion Quarterly* 29: 638–648.

Sudman, S. 1967. *Reducing the Cost of Surveys*. Chicago: Aldine.

Sudman, S. 1978. "Optimum Cluster Designs Within a Primary Unit Using Combined Telephone Screening and Face-to-Face Interviewing." *Journal of the American Statistical Association* 73: 300–304. Doi: http://dx.doi.org/10.2307/2286656.

United States Census Bureau. 2013. "Metropolitan and Micropolitan Statistical Areas Main." Available at: http://www.census.gov/population/metro/. Last Revised: May 6, 2013; (accessed January 15, 2015).

Wagner, J., K. Olson, and M. Edgar. 2013. *Using GPS and Other Data to Assess Errors in Level-of-Effort Data in Field Surveys*. Paper presented at the Joint Statistical Meetings, August 2013. Montreal, Canada.

Wang, K. and P. Biemer. 2010. *The Accuracy of Interview Paradata: Results from a Field Investigation*. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Chicago, IL. Available at: https://www.rti.org/sites/default/files/resources/aapor10_biemer_pres.pdf (accessed October 14, 2017).

Weeks, M.F., R.A. Kulka, J.T. Lessler, and R.W. Whitmore. 1983. "Personal versus Telephone Surveys for Collecting Household Health Data at the Local Level." *American Journal of Public Health* 73: 1389–1394.

West, B.T. and K. Olson. 2010. "How Much of Interviewer Variance is Really Nonresponse Error Variance?" *Public Opinion Quarterly* 74(5): 1004–1026. Doi: http://dx.doi.org/10.1093/poq/nfq061.

West, B.T., K.B. Welch, and A.T. Galecki. 2015. *Linear Mixed Models: A Practical Guide Using Statistical Software* (Second ed.). Boca Raton, FL: CRC Press.

Received February 2014
Revised July 2017
Accepted November 2017