Assessing Contact History Paradata Quality Across Several Federal Surveys

Nancy Bates[1], James Dahlhamer[2], Polly Phipps[3], Adam Safir[3], and Lucilla Tan[3]

[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2] National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782
[3]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212

Abstract
In 2004, the U.S. Census Bureau introduced an automated instrument to collect contact history paradata in personal-visit surveys. Survey methodologists analyze these data to improve, manage and evaluate surveys, for example, to plan contact strategies, predict survey nonresponse and assess nonresponse bias. But while the paradata literature is growing, a critical question remains - how accurate are the paradata themselves? We address this question by analyzing contact history data collected by the same instrument across three Federal surveys. We compare indicators of data quality to assess level of consistency both across and within the surveys. We also assess the degree of agreement between automated contact history data and information entered directly by the interviewer, such as attempt day and time and assessments of respondent cooperation.

KEYWORDS: Data quality, paradata, personal visit surveys

1. Introduction

Survey process data have always been critical to the measurement of survey quality (Couper and Lyberg, 2005; Scheuren, 2005; Lyberg, 2009). With the development of computer-assisted modes of data collection, data on the survey process automatically generated by the new electronic modes became known as "paradata" (Couper, 1998). The definition and scope of paradata now includes computer-generated as well as other types of data about the process of collecting survey data, specifically data that are not part of the survey interview (Kreuter and Casas-Cordero, 2010; Lynn and Nicolaas, 2010). Types of paradata include: call records generated electronically or by an interviewer, observations of interviewers and respondents, audio recordings of interviewer and respondent interactions, as well as items generated by computer-assisted instruments, such as response times and key strokes (Kreuter and Casas-Cordero, 2010).

Paradata are used to measure survey quality in a production environment, and to manage production with the goal of optimizing quality and minimizing costs (Couper, 2009). Fieldwork monitoring (Mockovak and Powers, 2008), non-response analysis (Bates, Dahlhamer, and Singer, 2008), and responsive designs (Groves and Heeringa, 2006) are all examples of the growing utility of paradata in survey operations. Paradata also has potential to aid in assessing measurement error (Kreuter and Casas-Cordero, 2010; Lynn and Nicolaas, 2010), for use in nonresponse adjustment (Maitland, Casas-Cordero, and Kreuter, 2009), and to improve editing and coding (Lynn and Nicolaas, 2010).

[1] Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau, the National Center for Health Statistics, or the Bureau of Labor Statistics.

While the uses of paradata continue to expand, a critical question remains – how accurate are the paradata themselves? Researchers have voiced this concern (Bates et al., 2008; Lynn and Nicolaas, 2010); however, few studies addressed this important question. One such study (Wang and Biemer, 2010), indicates the importance of studying paradata quality, suggesting that interviewer-generated call attempt data may be subject to underreporting.

In this study, our objective is to examine the quality of case or micro-level paradata in three federal surveys. Examining paradata quality is not an easy endeavor, as in this and most studies, there is no gold standard "truth" yardstick against which to measure the quality of paradata. However, we have the advantage of analyzing measures on paradata that have been collected in the same way across these surveys by one data collection organization. To examine the quality of these paradata, we assess measures of timeliness, consistency, and accuracy of contact attempt reporting by interviewers across the three surveys.

2. Background

The paradata focus of this paper is mostly on interviewer observations. The instrument used to record these observations is an automated instrument referred to as the Contact History Instrument (CHI). CHI provides interviewer-household observations for each contact attempt for all sample units, regardless of whether contact is made. Every time the survey questionnaire is accessed on the laptop, CHI launches automatically upon exiting the questionnaire, at which point, interviewers are expected to complete a CHI entry. Alternatively, a contact attempt entry can also be recorded by selecting a case from the Case Management System (CMS) and bringing up CHI without opening the survey itself. Interviewers can make a CHI entry immediately after a contact attempt or at a later time (for example, while in their car or at home). Interviewers are instructed to complete a CHI record each time a contact attempt is made.[2] Interviewer training for CHI consists of classroom and self-study and uses the same generic modules across all three surveys.

In addition to basic information such as date and time and mode of attempt, interviewers report the outcome of the attempt (e.g., contact with sample unit member, noncontact) and strategies employed before, during, or immediately after the attempt (e.g., left an appointment card, checked with neighbors, left promotional packet). For attempts resulting in a contact, interviewers complete a screen with 21 categories of verbal and nonverbal concerns and behaviors that may be expressed during interviewer-respondent interactions. Examples include "privacy concerns," "anti-government concerns," "too busy," and "hangs ups-slams door." Other screens collect information as to why an interview did not occur upon making contact (e.g., inconvenient time, respondent is reluctant, language barrier). Most CHI entry screens allow a "mark all that apply" format.

CHI data have been used previously to study a variety of topics including reasons for survey nonresponse, item nonresponse, contact patterns and strategies, nonresponse bias, and attrition (Maitland et al., 2009; Dahlhamer and Simile, 2009; Dixon, 2009; Bates et

[2]In theory, interviewers are expected to record a CHI entry whenever CHI automatically launches. However, the first CHI screen does have an "out" by allowing interviewers to select the category "Looking at a case – exit CHI". Therefore it is possible for interviewers to complete an interview without ever having recorded a single CHI entry.

al., 2008; Henly and Bates, 2006; Dahlhamer, Simile, Stussman and Taylor, 2005). Currently, CHI is the means for collecting automated contact histories in three ongoing surveys used to produce official statistics. These surveys are the source of data for our study.

The U.S. Census Bureau is the data collection agent for these three surveys. The primary mode of first interview is by computer assisted personal interviewing (CAPI). The first of these surveys to use CHI was the National Health Interview Survey (NHIS) beginning in 2004. The NHIS is an annual survey of the health of the civilian, noninstitutionalized household population of the United States, and is conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHIS produces nationally representative data on health insurance coverage, health care access and utilization, health status, health behaviors, and other health-related topics. Over 700 interviewers with the U. S. Census Bureau conduct the in-person interviews (some telephone follow-up is allowed). Case assignments are released to interviewers each week throughout the calendar year (except the first two weeks of January) and are to be completed within a 17-day interview period. Each year interviews are conducted in approximately 35,000 households yielding data on roughly 87,500 persons. All NHIS analyses presented in this paper are based on case-level data (n=64,540) and on attempt-level CHI data (n=246,718) collected during the 2009 calendar year.

The second survey to use CHI was the Consumer Expenditure Interview Survey (CE) beginning in April 2005. The CE is sponsored by the Bureau of Labor Statistics and collects data on spending by America's consumers. The CE is a household panel survey conducted over five consecutive calendar quarters. Each of the five quarterly interviews is referred to as a "wave" of data collection, with an average of 7,000 completed interviews per wave. Over 700 interviewers with the U.S. Census Bureau conduct the in-person interviews (some telephone follow-up is allowed). For each month's sample, interviewers receive their assignments prior to the start of the month, begin interviewing on the first day of the month, and end interviewing prior to the start of the next month. All CE analyses presented in this paper are based on Wave 1 household survey data (n=12,106) and attempt-level CHI data (n=47,652), collected during the 2009 calendar year. We chose to subset the CE analysis to Wave 1 data for two reasons: first, for comparability, since the NHIS is a one-time survey, and second, because restricting the data to first interviews allows an analysis of contact attempt data uncontaminated by respondent experiences from prior interviews.

The third survey included in our study is the Current Population Survey (CPS), a monthly household survey sponsored by Bureau of Labor Statistics. The CPS provides comprehensive data on the labor force, employment, unemployment, persons not in the labor force, hours of work, earnings, and other demographic and labor force characteristics. The CPS has a panel design in which selected households are included in the survey for a total of eight monthly interviews, using a rotation schedule (households are in the survey for four consecutive months, out for eight months, and back in the survey for another four consecutive months). Each month approximately 54,000 interviews are completed. CHI was introduced to the CPS in July 2009. Again, for purposes of comparability across surveys, we limit our analysis to households in their first CPS monthly interview (wave 1). Because the CHI was only introduced in July 2009, we limited our analyses to cases from August to December 2009. This represents 109,362 attempt-level CHI records collected for 44,897 CPS cases. Census interviewers collect the first month of CPS data via in-person visits (some telephone follow-up is

allowed). As a monthly survey, the CPS has a relatively short interviewing assignment period of nine days. Interviewers receive their case assignments on a Tuesday, begin interviewing on Sunday, and end interviewing a week later on the following Monday.

3. Description of Contact Attempt Efforts

Before proceeding to the more substantive analysis of the quality of CHI data, we first examined descriptive statistics for CHI across the three surveys, including contact attempt characteristics (section 3.1) and patterns of interviewer use (section 3.2). The purpose of this effort was to gain insight to overall trends on various characteristics of the data collection effort for each survey. The contact attempt characteristics included distribution of first contact attempts by day and time, number of contact attempts needed to reach final disposition, number of respondent concerns reported, number of interviewer strategies reported, and number of contact attempts by contact attempt outcome. The patterns of interviewer use measures included how interviewers accessed CHI, time needed to enter CHI data, and non-use of CHI. We examined descriptive statistics across the full contact history (not just one attempt) in order to develop a clearer picture of contact attempt patterns, as recorded by interviewers, as well as to gain a better understanding of how interviewers actually use CHI. This information is helpful for the later interpretation of our data quality tabulations, and it provides an illustration of how CHI can be used both in the field and for analysis.

Additionally, since one of the main objectives of the analysis is to identify variations in CHI reporting that might be suggestive of data quality concerns (whether observed in timeliness, consistency, or misreporting), focusing first on contact attempt characteristics and patterns of interviewer use provides an early indication of the extent to which differences in CHI are due to fundamental survey design characteristics or field practices, or possibly some other source. In other words, can data quality differences, if observed, be explained by basic differences in contact attempt patterns or interviewer use? The answer to this question will be useful for guiding management or field efforts to address potential deficiencies in the quality of CHI data.

3.1 Characteristics of Contact Attempts

3.1.1 Day and Time of Contact Attempt

We examined the distribution and outcome of first contact attempts by day and time, to learn more about when first contacts are attempted, as well as their relative success. As seen in Table 1, the majority of first attempts for the three surveys occurred Monday through Thursday, during the day,[3] closely followed by Monday through Thursday, during the evening. The three surveys also exhibited similar first attempt patterns during both weekday days (38.4 percent for CE, 39 percent for CPS, and 51.3 percent for NHIS[4]) and evenings (27.3 percent for CE, 32 percent for CPS, and 32.7 percent for NHIS). As has been documented in other research on household surveys, weekdays are not as productive relative to other day and time periods (e.g., see Carley-Baxter, Peytchev, and Black, 2010; and Massey, Wolter, Wan, and Liu, 1996; Weeks, Kulka, and

[3] Time slots include day (8:00am until 4:00pm), evening (4:00pm until 9:00pm), and overnight (9:00pm until 8:00am).

[4] The estimate for NHIS daytime first attempts may be higher than that of the other surveys due to a requirement for its interviewers to always start interviewing for a new field period on a Monday, whereas CPS always starts its new field period interviewing on a Sunday, and CE starts on a variable day of the week depending on what day the first of the month occurs.

Pierson, 1987). In fact, in the three surveys included in this analysis, while more first contacts were attempted during the week's daytime hours, it was the evening hours that were more productive in terms of resulting in a contact (30.9 versus 40.8 percent for CE, 34.5 versus 41.1 percent for NHIS, and 38 versus 50 percent for CPS).

3.1.2 Number of Contact Attempts to Final Disposition

We reviewed the number of contact attempts needed to reach final disposition for the three surveys, and found general agreement across the three surveys (see Table 2): *noninterview-no one home* cases had the highest mean number of contact attempts (5.3 for CPS, 8 for CE, and 8.5 for NHIS), while *out-of-scope* cases had the lowest mean number of contact attempts (1.8 for CPS, 2.4 for CE, and 2.6 for NHIS). It is worth noting that for both these final dispositions, CPS had a lower number of contact attempts. This difference may be due to distinct survey design characteristics, such as length of field period (CPS is in the field for just 9 days, NHIS is in the field for 17 days, and CE has a one-month field period), response rate differences[5] (CPS has an average response rate of 92 percent, NHIS response rate is 82 percent while CE has an average response rate of 75 percent), or the relative newness of CHI for CPS field representatives. Nevertheless, this is an important phenomenon to understand, as contact attempts are costly and should be minimized when justified.

3.1.3 Respondent Concerns

The next measure we examined was the number of respondent concerns reported. Since interviews are likely to be completed among more cooperative respondents, we expected a high prevalence of no concerns reported for *completed interviews*; but this prevalence was less than 75 percent of completed interviews in all three surveys (56 percent for NHIS, 58.6 percent for CE, and 70 percent for CPS). However consistent with expectations, the prevalence of no concerns reported among *refusal* cases was relatively low in the three surveys (4.2 percent for CE, 5.8 percent of NHIS, and 6.3 percent for CPS). These rates suggest that interviewers are collecting information on survey participation concerns on approximately 95 percent of *refusal* cases.

3.1.4 Strategies to Gain Respondent Cooperation

Next we reviewed the number of respondent cooperation strategies reported by interviewers, including contact attempt and gaining cooperation strategies such as providing an advance letter, promotional packet, or informational brochure; scheduling a follow-up appointment; and checking with neighbors. On this measure, the three surveys were in general agreement: 86.6 percent of NHIS, 86.1 percent of CE, and 80 percent of CPS contact attempts had at least one strategy reported. Looking at the data by case disposition, *refusals* showed a higher percentage of at least one strategy reported (98.7 percent for CE, 90.8 for NHIS, and 80.2 for CPS) as compared to *completed interviews* (84.4 percent for NHIS, 82.3 for CE, and 80.1 for CPS), which is to be expected.

3.1.5 Contact Attempts by Contact Outcome

The final usage measure we examined was the proportion of contact attempts resulting in contact with a sample unit member (as opposed to contact with a non-sample unit member or a noncontact). Consistent with prior results, the surveys exhibited similar rates: 57 percent of NHIS, 58 percent of CE, and 66 percent of CPS contact attempts resulted in contact with a sample unit member.

[5] The NHIS response rate is equivalent to AAPOR RR6. For CE it is equivalent to AAPOR RR1 and for CPS it is RR2 (AAPOR, 2009).

3.2. CHI Use by Interviewers

In this section, we describe how interviewers access CHI, the amount of time they spend recording contact attempt histories, and the degree to which they neglect to make any CHI entries altogether. Interviewers have two ways to make a CHI entry – "actively" via selecting a case from the CMS without opening the survey instrument, or "passively" when the survey instrument is open and CHI launches automatically upon exiting the survey. Most CHI entries were made "passively" in all three surveys: 80.4 percent of CHI entries for CE, 88.5 percent for NHIS, and 96 percent for CPS. One explanation for CPS having a lower proportion of "active" CHI entries via CMS may simply be a function of the relatively short CPS field period. Interviewers have less time to manage and work their caseloads, and thus less time to perform auxiliary tasks in addition to basic survey data collection.

We also looked at how much time interviewers were taking to record the contact history data using CHI timer data on each case. The median time taken to record each contact attempt was under one minute for all three surveys: 43 seconds for NHIS, 45 seconds for CPS, and 55 seconds for CE.[6] By final disposition, *refusals* tended to have a longer median time with 58, 70, and 77 seconds for the NHIS, CPS, and CE, respectively, while *completed interviews* have the lowest median time (41 seconds for CPS, 42 seconds for the NHIS, and 49 seconds for CE).[7] This is not surprising given that the typical CHI path to record, for example, an attempt resulting in a noncontact consists of only 6 screens, each requiring a minimum of a single radio button click and taking only a few seconds from beginning to end. Additionally, the surveys only averaged around 2 to 4 contact attempts recorded per case. These numbers suggest that recording contact histories in an automated environment can be streamlined such that the interviewer burden is light – perhaps a fact that can be emphasized during training or when a survey program is considering implementing CHI.

We next attempted to quantify the extent of missing CHI data by seeing how often interviewers neglected to record any contact history for cases. As with many new interviewer procedures, we expected a learning curve (and perhaps even some resistance) to the use of CHI. We hypothesized this might also vary across the surveys since they differed in the length of experience using CHI (only 6 months for CPS compared to 4 years for CE and 5 years for NHIS). We found that the proportion of cases without any CHI data was very low for NHIS and CE, but slightly higher for CPS (1.1 percent for NHIS, 1.6 percent for CE, and 4.5 percent for CPS). The higher incidence in the CPS is presumably because interviewers were less acclimated to it at the time of our study. Nonetheless, the numbers are reassuring and suggest that most of the time, the surveys were capturing some minimal amount of contact history information.

We did not find any consistency in cases missing contact histories by final disposition. The prevalence was highest among *out-of-scope* cases for CE, among c*ompleted interviews* for CPS, and among *noninterview-language problem* cases for NHIS. We also examined differences in the degree of missing data across the Census regional offices (RO). The ROs consist of 12 decentralized offices that are to some degree independent,

[6] We elected to report median time rather than mean time as the means were influenced by severe outliers, presumably cases where interviewers may have left the CHI instrument open for long periods of time.

[7] Time spent recording CHI entries for out of scope cases in the NHIS were practically identical at 41 seconds.

each with its own unique set of procedures and culture. Debriefings held with RO staff when CHI was first released suggested that the perceived utility and adoption of CHI varied across the ROs (Ruffin, J. 2006). We found some noticeable variation among the ROs, with three ROs in particular exhibiting higher rates of missing CHI data.

Next, we examined the prevalence of cases missing CHI data among interviewers. We wondered if the distribution of cases without CHI data was concentrated among a few interviewers. The interviewer workloads varied substantially among the three surveys (median workload of 15 cases for CE, 29 for CPS, and 61 for NHIS).[8] To account for the different workloads among interviewers, we examined this measure for the subset of interviewers with workloads of at least 10 cases, and the proportion of those interviewers with 25 percent or more of their cases without CHI data. We found that only a handful of interviewers met these criteria: around 1 percent (6 interviewers) for NHIS, 2 percent (8 interviewers) for CE, and 6 percent (71 interviewers) for CPS (see Table 3). It is reassuring that the overwhelming majority of interviewers are recording some amount of contact history information for most of their caseload. Still, this information could be used by supervisors and management to identify ROs and specific interviewers to better understand the reasons and circumstances for neglecting to record paradata. If necessary, supervisors could consider retraining some interviewers on the use of CHI.

With an understanding of how CHI is used in the field, we next turn to examining the quality of CHI data in terms of timeliness, consistency, and errors in reporting.

4. CHI Data Quality Indicators

4.1. Timeliness

CHI was built with the recognition that some situations might not lend themselves to an interviewer recording contact attempt details immediately after the attempt. For example, during a CE focus group with interviewers, it was suggested that when "running down a case," driving by a sample unit address, or making repeated calls to the same case, interviewers might record CHI entries after the fact or not at all (Edgar, 2006). We hypothesize that the sooner in time the interviewer records a CHI entry after a contact attempt, the less likely the entry is subject to recall error. We examined the distribution of the timing of CHI entries – this is captured by a screen in CHI that specifically probes for this information. We found that most CHI entries were recorded immediately after the actual contact attempt (81 percent for CE, 90 percent for NHIS, and 92 percent for CPS), so CHI data are being logged in a fairly timely manner and are less likely to be subject to serious recall error (although we note the CE, with close to 20 percent of its contact histories recorded after-the-fact may be cause for further investigation). We also found that the recording of CHI entries was more likely to be postponed for contact attempts that resulted in *noncontact* (11 percent for CPS, 13 percent for NHIS, and 23 percent for CE) than for those resulting in *contact* (5 percent for CPS, 6 percent for NHIS, and 13 percent for CE). Our hypothesis is that interviewers are less likely to open the instrument when attempts result in noncontacts, thus missing the automatic prompt of the CHI screen, so they have to remember to make the CHI entry at a later time.

[8] Interviewer workloads are based on cases in the current study: 2009 wave 1 cases for CE, 2009 cases for NHIS, and 5 months (August to December) of CPS month-in-sample 1 cases in 2009.

4.2. Consistency of Reporting

Interview and process data collected in the survey instruments (e.g., final case dispositions, dominant mode of data collection, cooperativeness of respondents, responses to sensitive items), the majority of which are interviewer reported, provide unique opportunities to examine consistency across the survey and CHI instruments. We started by examining various final noninterview dispositions, and checked for related CHI entries. For cases with a final disposition of *noninterview-language problem*, we checked for how many of these cases were coded in CHI as unable to conduct the interview due to a language problem (at 1 or more contact attempts). Moderate variation was observed across the three surveys with consistency rates ranging from 45 percent for CE, to 65 percent for CPS, and 78 percent for NHIS.

Next we explored cases with a final disposition of *noninterview-no one home* with at least one personal visit attempt (as recorded in CHI), identifying the percentage where at least one CHI entry of "no one home," "no one home—appointment broken," or "no one home—previous letter/note taken" was recorded. Consistent reporting was evident in all three surveys. In over 90 percent of these cases, one or more of the corresponding categories were reported in CHI. We also examined *no one home* cases with at least one telephone attempt (as recorded in CHI). For telephone-based noncontacts, interviewers can record a range of reasons in CHI including "got answering machine/service" and "no answer." While the rate of consistent reporting between these two CHI entries and the final case disposition of *noninterview-one home* was lower than that observed for personal visit attempts, the rates were again consistent across the three surveys (68 percent for CE, 70 percent for CPS, and 72 percent for NHIS).

We also examined final *noninterview-temporarily absent* disposition cases with at least one personal visit attempt. One related reason for a personal visit noncontact available in CHI is "on vacation, away from home/at second home." Here we observed much lower rates of agreement: 31 percent for CPS, 32 percent for CE, and 38 percent for NHIS. Given its wording, many interviewers may construe this CHI category to be limited to residents away for leisure-based reasons. Hence, interviewers may be reluctant to include other reasons such as "in the hospital" or "away on a business trip". Revision of the existing category and/or the addition of new categories may be warranted.

An important use of CHI data is to identify less cooperative cases early in the field period, enabling any number of field actions such as transferring these cases to more experienced interviewers or refusal converters, automatically switching these cases to telephone administration to reduce costs, and/or using more targeted recruitment protocols such as financial incentives. For these strategies to be effective, interviewers must accurately record the outcomes of each attempt including any concerns or reluctance expressed by householders. As a check, we limited our analysis to cases with a final *refusal* disposition and examined the percentage of such cases that were recorded as a "soft refusal" (i.e., "respondent is reluctant") in CHI. The rate of agreement was lower than anticipated: 50 percent for CPS, 59 percent for CE, and 64 percent for NHIS.

Our consistency checks on final case dispositions and related entries in CHI focused on *completed interviews* and *out-of-scope* cases. First, we examined the percentage of *completed interviews* where the interviewer failed to record contact with a sample unit member in CHI. Inconsistent reporting in this instance was quite rare: less than 1 percent

for CE and NHIS, and 1.3 percent for CPS. We also looked at final *out-of-scope* disposition cases with at least one personal visit attempt (as recorded in CHI), and the congruent reporting of "completed case, out-of-scope" as a reason for personal visit noncontact in CHI. We found a fairly low rate of consistent reporting, and considerable variation across the surveys: 43 percent for NHIS, 65 percent for CPS, and 70 percent for CE.[9]

The next set of consistency checks we performed focused on the mode of contact attempts recorded in CHI and the interview mode. Questions on mode of administration were generally captured at the end of the survey instruments, with the questions varying across surveys. Regardless, rates of inconsistent reporting were quite low in all three surveys. In the NHIS, interviewers are asked if any main modules (household composition, family, sample child, sample adult) of the survey instrument were completed primarily by telephone. For roughly 1 percent of NHIS interviews meeting this criterion, no telephone contact attempts were recorded in CHI. Conversely, for just under 1 percent of interviewed cases where none of the main modules was completed primarily by telephone, no personal visit attempts were recorded in CHI. For the CPS, interviewers are asked if most of the data were collected by telephone. In 2.5 percent of interviewed cases where this was true, no telephone contact attempts were recorded in CHI. For interviewed cases where most of the data were collected by personal visit, only 0.3 percent were lacking personal visit entries in CHI. For the CE (beginning in April 2009), interviewers are asked if all sections of the survey are collected by personal visit, all sections by phone, or a mix of phone and personal visit. Among interviewed cases where all contact attempts were recorded as personal visits in the survey instrument, only 0.1 percent did not have a personal visit attempt recorded in CHI. Conversely, of those interviewed cases where all contact attempts were recorded as telephone attempts in the survey instrument, 3.1 percent did not have a single phone attempt recorded in CHI.

A final consistency check involving the survey instrument and CHI looked at the reporting of "privacy concerns" in CHI by responses to income questions in the CPS and NHIS.[10] The NHIS asks a single, exact amount question on total family income for the previous calendar year. As expected, the percentage of cases where "privacy concerns" were reported in CHI was considerably higher when refusal responses were given to the income question (36.7 percent), as opposed to don't know responses (15.2 percent), and exact amounts (10.1 percent). Similar patterns were observed for the CPS, which, in the first interview, asks a single, categorical question on total family income for the past 12 months. In nearly 24 percent of cases where total family income was refused "privacy concerns" were recorded in CHI. Only 8.4 percent of "don't know" cases and 5.9 percent of income reporters had CHI entries of "privacy concerns."

We also checked for consistencies within CHI with a focus on survey mode. We were interested in the extent to which interviewers recorded the use of telephone-based strategies for making contact or securing participation during personal visit attempts, and vice versa. We found that overall inconsistencies in reporting were fairly low for all three surveys, but slightly higher than that observed for some of the previous mode-based analysis. Among personal visit attempts, telephone-based strategies, including "called

[9] A large percentage of out-of-scope cases in the NHIS results from households "screened out of the survey" on the basis of race/ethnicity. A much smaller number of out-of-scope cases results from households occupied entirely by Armed Forces members, minors, or persons with a usual residence elsewhere. If we exclude these "screened out" cases from the calculation, the rate of agreement improves to 63 percent.
[10] The CE does not ask income questions in the wave 1 interview.

household," "left message on answering machine," and "called contact persons," were reported for 2 percent for NHIS, 3.5 percent for CE, and nearly 4 percent of personal visit attempts for CPS. A higher rate of inconsistent reporting was observed when focusing on telephone attempts. Personal visit strategies, including "advance letter given," "left note/appointment card," "left promotional packet/informational brochure," "stake-out," and "checked with neighbors," were reported for 5 percent of telephone attempts for NHIS, 6 percent for CE, and 8 percent for CPS. One possible explanation for these discrepancies may be the cumulative recording of strategies. That is, for any given contact attempt interviewers may record all strategies used to that point.[11] It is also possible that what appear to be telephone- or personal visit-specific strategies are not always interpreted as such by interviewers. Focusing on the NHIS, in nearly 3 percent of telephone contact attempts interviewers reported giving an advance letter or leaving a note/appointment card. It is quite plausible that an interviewer could read the advance letter over the phone or consider "leaving a note" a strategy to be checked in conjunction with "leaving a message on an answering machine." Finally, during focus groups conducted in 2006, CE interviewers indicated that both personal visit and telephone strategies are sometimes employed on the same contact attempt -- for example, when an interviewer attempts to contact a respondent by telephone during a failed personal visit contact attempt. In such situations, interviewers must choose to report only one mode for the contact attempt, leading to the possible appearance of inconsistency in their reporting of strategies used.

4.3. Mis-reporting in CHI

Reporting illogical or non-applicable responses within a CHI screen are other indicators of poor data quality that we examined. For example, for the question on the screen that allows interviewers to pick from a long list of concerns that a contact sample unit member may express about survey participation, there is also a "no concerns" category. We looked at the number of times the interviewer checked off one or more "concern" categories as well as the "no concerns" category. For all three surveys, this occurrence was extremely rare with around only 1 percent of CHI records indicating this type of mis-reporting.

In addition, we also examined the prevalence of non-applicable categories reported in CHI entries. For the CHI screen on concerns about survey participation, the selection of longitudinal-type categories[12] (e.g. "gave same information last time," "intends to quit survey," "requests same interviewer as last time") would not be applicable; for the CHI screen on strategies, the selection of survey-specific categories would not be applicable (e.g., "CED double placement" applies only to the Consumer Expenditure Diary survey). We found these types of mis-reporting to also be extremely rare for all three surveys as they occurred in less than 1 percent of all CHI entries. We did find that 2.3 percent of the CPS cases had a code of "respondent requests same interviewer", however, this could simply indicate respondents requesting the same interviewer for *future* interviews (CPS cases are in sample for an additional seven interviews after the first interview).

[11] Of course, we might expect the percentage of discrepancies to be consistent across mode of attempt.
[12] Recall for this analysis, our universe of study is restricted to one time interviews (NHIS) or time-in-sample 1 cases (CPS and CE).

5. Summary and Conclusions

Findings from our study provide information on a topic that has received little to no attention in the survey methods literature -- namely, quality of paradata. Our study capitalized on our access to three different personal-visit surveys, all using the same automated instrument to record interviewer observation paradata, thus providing identical metrics to evaluate. We attempted to address three questions. First, what can we learn about the quality of the paradata? Second, what does this tell us about using these paradata to make survey management decisions in the field? Finally, how might we improve the paradata collection instrument itself?

Using previous literature as a guide, our findings suggest that our day and time of attempt paradata are likely an accurate reflection of when interviewers are making their contact attempts. Similar to other studies we found that interviewers make most of their attempts during weekdays during the day (between 9am and 4 pm) but are more successful making contact during weekday evenings. CHI data also indicate that *no one home* cases require the most contact attempts and *out of scope* cases the least – findings also backed up by previous studies. The fact that these findings were replicated across the different surveys also lends credence to the reliability of the measure.

We also saw evidence that, by and large, interviewers are dutifully recording at least some of their interviewer observations. The prevalence of contact attempt history records completely missing for cases was low (only between 1-5 percent of all cases), and this occurrence was concentrated among a few interviewers. We view this as positive news and due in part perhaps to the fact that the observations can be recorded quickly (most interviewers spent less than one minute total recording their observations per case). On the other hand, compliance may also come from the fact that interviewers are confronted with CHI every time they exit the survey instrument, and in fact, the overwhelming majority of attempts are being recorded via this passive method as opposed to actively recording histories via the CMS. While nearly all cases had at least one paradata record, our findings cannot confirm if interviewers accurately report their number of contact attempts. We did confirm that, on average, between 2 and 4 attempts are being recorded depending upon the survey, but interviewer survey results suggest that CE interviewers estimate they only complete a record for around 85 percent of their attempts, which is of concern (Mockovak, Edgar, and To, 2010). Situations described by the interviewers that contribute to underreporting include: multiple drive-by attempts, when using cell-phone for contacts, when contact is with non-sample unit member, when setting appointments before the start of the interview field period, and after 7 or more attempts have been made.

Other data quality indicators suggest that interviewers are being conscientious when describing their contact attempts. For example, it was a rare occurrence for interviewers to report a strategy that was inconsistent with the mode identified and even rarer when they recoded implausible pairs of codes or codes that did not apply to their particular survey. Similarly, the mode indicator as measured by CHI was found to be extremely consistent with the mode indicator as recorded in the survey instrument. We also learned that the majority of contact histories are recorded immediately after the attempt versus at a later time – good news when considering recall bias as a source of poor data quality.

This is not to say the data are without quality or consistency problems. When checking to see how often interviewers recorded CHI entries that paralleled the final disposition code

of the case (e.g., language problem, soft refusals, no one home), the consistency was not as high as we expected, and this could have important implications for the field. For example, we found that for cases with a final disposition of *refused*, in most cases (around 95 percent) interviewers recorded at least one specific type of concern at the appropriate screen. However, far fewer interviewers identified these refusers as "respondent is reluctant" when asked to indicate the reason for making a contact without an interview. On this indicator, interviewers were consistent only 55 percent to 70 percent of the time (depending upon the survey). To feel confident using these data as a predictor of final outcomes, we would prefer more agreement. Further multivariate analysis is recommended in this area whereby interim outcomes recorded in the paradata could be used to model and predict final outcomes. This is critical to understanding what paradata may be most (and least) useful in responsive designs and other real-time field interventions.

Overall, our findings suggest that analysts can place confidence in the paradata currently being collected across the three surveys. Used with some care, the metrics provide a meaningful narrative of the events leading up to final outcomes. This includes the causes for respondent concern, the different strategies and modes employed, and the time and day pattern of attempts. Each of these is useful when making decisions such as interviewer re-assignments, nonresponse conversion plans, and/or targeted appeals. However, there is also room for improvement. Although we did not explicitly measure it, we believe some degree of underreporting of contact histories is occurring. One clue is that attempts resulting in a noncontact are more likely to be recorded later as opposed to at the time of attempt. It is plausible that some noncontacts are never recorded. We also saw in several instances that the survey with the least amount of experience collecting paradata (CPS) may have lower data quality (i.e., more missing paradata spread among a greater number of interviewers, fewer histories recorded on average, and less consistency between observations and expected final outcomes). As interviewers become more acclimated to recording the paradata, these levels will likely improve.

In terms of recommendations for improving the contact history instrument itself, we offer several thoughts. First, although extremely rare, we did see a few instances where an illogical pair of interviewer observations was recorded (e.g., where interviewer marked "no concerns" but also indicated the respondent had "privacy concerns"). This can be remedied simply by re-programming the instrument to not allow additional entries if the "no concerns" category is selected. This change was recently implemented. We also saw instances where category wordings were probably interpreted differently by interviewers, either across surveys (e.g., on vacation, away from home/at second home may be interpreted as applicable to households temporarily absent for only leisure activities), or within a survey (e.g., CPS interviewers, perhaps recording future requests by selecting "requests same interviewer as last time" as a survey participation concern during the first interview). These concerns could be remedied by making changes to the category wording, by programming edits type checks, or through additional training.

Finally, field supervisors and interviewers can be and have been asked to provide feedback to improve CHI. A survey of CE interviewers identified a number of issues (e.g., instrument is slow when opening CHI, and limitations and inconsistencies in recording attempts when the contact is a non-household member (Mockovak, Edgar, and To, 2010). In some cases, interviewers have developed work-around solutions for inconsistencies; however, permanent solutions will be necessary to implement to assure optimal data quality. Since several years have passed since interviewers were last queried

about CHI, and because an additional survey is now using it, we recommend a new round of interviewer and management assessments. CHI has been in use at Census for six years and is probably due for a comprehensive review.

Regarding further research, we make several suggestions. First, we suggest looking to new technologies to further assess paradata validity and quality. If possible, the use of computer-assisted recorded interviewing (CARI) might be implemented. Ideally, we could record the pre-interview door-step interactions so we could have the "truth" against which to compare CHI entries. However, given the legal and policy requirements to obtain informed consent prior to using CARI, this may prove impossible. An alternative is to have trained observers shadow interviewers, record their own versions of CHI, and then compare their records and the interviewer's. Second, we recommend bringing interviewer characteristics into the equation when assessing paradata quality (e.g., years of experience, gender, education). Since recording interviewer-respondent interactions is a rather subjective undertaking, interviewers are undoubtedly a source of systematic variance. To date, there is very little research regarding interviewer impact on the collection of paradata.

Acknowledgements

We owe special thanks to Dori Allard, Lisa Clements, Marcie Cynamon, Jennifer Edgar, Jane Gentleman, Matt Jans, David Sheldon, and Elizabeth Sweet for their helpful comments on this paper.

References

American Association for Public Opinion Research. (2009). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 6th edition*. AAPOR.

Bates, N., J. Dahlhamer, and E. Singer (2008). "Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse." *Journal of Official Statistics* 24 (4): 591-612.

Carley-Baxter, L., Peytchev, Andy, and Black, M.C. (2010). "Comparison of Cell Phone and Landline Surveys: A Design Perspective," *Field Methods*, Vol. 22, No. 1, 3-15.

Couper, M. (1998). Measuring survey quality in a CASIC environment. Pp. 41-46 in Proceedings of the Section on Survey Research Methods. Alexandria, VA: American

Couper, M. (2009). "The Role of Paradata in Measuring and Reducing Measurement Error in Surveys." Presented at the National Center for Social Research, Network for Methodological Innovation. London (August).

Couper, M. and L. Lyberg (2005). "The use of paradata in survey research." Proceedings of the 55th Session of the International Statistical Institute, Sydney, Australia.

Dahlhamer, James M., Catherine M. Simile, Barbara J. Stussman, and Beth Taylor (2005). "Determinants and Outcomes of Initial Contact in the National Health Interview Survey, 2004." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Miami Beach, FL, May 14.

Dahlhamer, James M. and Catherine M. Simile (2009). "Subunit Nonresponse in the National Health Interview Survey (NHIS): An Exploration Using Paradata." *Proceedings of the Joint Statistical Meetings*, Washington, DC.

Dixon, J. (2009). "Modeling the Difference in Interview Characteristics for Different Respondents." Proceedings of the Joint Statistical Meetings, Washington, DC.

Edgar, J. (2006). "FR Focus Group Results: CE Concepts". Internal Bureau of Labor Statistics memorandum, from Jennifer Edgar to BRPD, J. Ryan, C. Pickering S. Groves and B. Mockovak, September 8.

Groves, R. M. & Heeringa, S. (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: 439-457.

Henly, M. and N. Bates (2006). "Using Call Records to Understand Response in Longitudinal Surveys." Presented at the American Association for Public Opinion Research Annual Meeting, Montreal.

Kreuter, F. and C. Casas-Cordero (2010). Paradata. German Council for Social and Economic Data Working Paper Series, Berlin, Germany, No. 136 (April).

Maitland, A., Casas-Cordero, C. and Kreuter, F. (2009). An evaluation of nonresponse bias using paradata from a health survey. Pp. 2250-2255 in Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Lyberg, L. 2009. "The paradata concept in survey research." Presented at the National Center for Social Research, Network for Methodological Innovation. London (August).

Lynn, P. and G. Nicolaas. (2010). "Making Good Use of Survey Paradata." *Survey Practice*, April: www.surveypractice.org.

Massey, J. T., Wolter, C. Wan, S. C., and Liu, K. (1996), "Optimum Calling Patterns for Random Digit Dialed Telephone Surveys," Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: 485-490.

Mockovak, W., J. Edgar, and N. To. (2010). "Results from the CEQ Field Representatives Survey, Fall 2009." Internal Bureau of Labor Statistics report (June).

Mockovak, W. and R. Powers (2008). "The use of paradata for evaluating interviewer training and performance." Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Ruffin, J. (2006). "Contact History Instrument (CHI) Debriefing Session". U.S. Census Bureau internal memorandum, from Josephine Ruffin to Somonica Green, Chief, Financial Surveys Branch, December 18.

Scheuren, F. (2005). Paradata from concept to completion. Proceedings of the Statistics Canada International Symposium Series. Ottawa: Statistics Canada.

Wang, K. and P. Biemer (2010). "The Accuracy of Interview Paradata: Results from a Field Investigation." Presented at the Annual Meeting of the American Association for Public Opinion Research, Chicago, IL (May).

Weeks, M.F., R. Kulka, A. and Pierson, A. (1987). Optimal Call Scheduling for a Telephone Survey. Public Opinion Quarterly51 (4): 540-549.

Table 1. Day of Week and Time of 1st Contact Attempt & Outcome

	% Distribution of Day and Time of 1st Contact Attempts			% of Contact Attempts where Contact Made		
Day-Time	CE	CPS	NHIS	CE	CPS	NHIS
	n=12,106	n=44,894	n=64,410			
Overnight	2.3	2.6	1.8	24.8	30.5	27.9
Monday-Thursday Day	38.4	38.7	51.3	30.9	38.1	34.5
Monday-Thursday	27.3	32.3	32.7	40.8	49.6	41.1
Friday Day	7.5	1.4	3.8	31.8	42.3	36.2
Friday Evening	4.1	1.1	2.4	34.1	53.1	41.1
Saturday Day	9.0	1.2	4.0	38.0	58.2	46.3
Saturday Evening	4.0	.7	1.7	32.3	52.0	42.7
Sunday Day	4.1	12.8	1.2	32.9	45.6	45.5
Sunday Evening	3.3	9.3	1.0	37.7	43.2	41.1

Note: Day: 8:00<16:00, Evening: 16:00-<21:00, Overnight: 21:00-<8:00

Table 2. Mean Number of Contact Attempts by Final Disposition

Final Disposition of Case	CE	CPS	NHIS
Completed interview	3.6	2.3	3.9
Nonresponse			
Refusal	6.4	4.8	6.8
No one home	8.0	5.3	8.5
Temporarily absent	6.9	4.9	6.6
Language problem	7.1	3.7	5.6
Out of scope	2.4	1.8	2.6
Total	3.9	2.4	3.8

Table 3. Prevalence of Missing CHI Data

	CE	CPS	NHIS
Percent of cases without CHI records	1.6	4.5	1.1
Number of interviewers	718	1,371	765
Number of interviewers with workload of 10 or more cases	458	1,229	670
Percent of interviewers missing CHI Records for 25% or more of caseload	1.7	5.8	0.9