

STUDY SERIES
(*Survey Methodology #2011-03*)

**Requesting Consent to Link Survey Data
to Administrative Records: Results from
a Split-Ballot Experiment in the Survey
of Health Insurance and Program
Participation (SHIPP)**

Joanne Pascale

Center for Survey Measurement
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: June 22, 2011

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

**Requesting Consent to Link Survey Data to Administrative Records:
Results from a Split-Ballot Experiment in the Survey of Health Insurance and Program
Participation (SHIPP)**

Abstract: Administrative records have begun to play a key role in survey research and, while policies regarding consent are still in flux, there is general agreement that research is needed on how to communicate to respondents what linking entails, and how to motivate their consent. Previous research found that 26 percent of those initially opposed to data sharing shifted their position when prompted with arguments about potential improvements in accuracy and reductions in cost (Singer and Presser, 1996). In order to take these findings further, in the spring of 2010 a field experiment was carried out by the US Census Bureau which included three panels, each presenting a different rationale to the respondent for data linkage: improved accuracy, reduced costs, and reduced respondent burden. Somewhat contrary to expectations, there was no statistically significant difference in consent rates across the three versions of the consent question. Overall levels of consent, however, were rather high (84 percent), and represented a shift of more than 20 percentage points compared to a similar study in 2004. Demographic analysis indicated that age and non-response to a household income question were predictors of both levels of consent and missing data on key variables needed to make that linkage. Education and sex were also predictors of objection rates. There was also some evidence of interviewer effects; one of the three interviewer groups had a higher rate of respondent objections to consent and a higher rate of missing data on key fields of data used for record linkage.

Key Words: Implicit or Informed Consent, Administrative Record Linkage

INTRODUCTION

Administrative records play a role in survey research in a number of ways, such as validating survey reports, supplementing or even substituting for survey data. In large part due to declining response rates and increasing costs, as well as the persistent undercount of minority populations in the decennial census, researchers are looking to make increased use of records in data collection. Among the challenges in this area are understanding the public's knowledge of and attitudes toward data sharing in general, and exploring how to explain to respondents just what data sharing entails, and how to request and motivate their consent to link survey data to administrative records.

Existing research on this topic is fairly limited, but in response to concerns about data quality following the 1990 decennial census, the National Academy Panel on Census Methods recommended a research program to explore the use of administrative records in government. Among the components of research carried out was a computer-assisted telephone survey about confidentiality and data sharing with regard to the decennial census in particular, and attitudes toward government and privacy in general. The survey also examined whether those initially opposed to data sharing for the decennial census would shift their position when prompted with arguments about potential improvements in accuracy and reductions in cost. These results were fairly positive, showing that 26 percent were persuaded to a more favorable position in light of these arguments (Singer and Presser, 1996). One area not explicitly addressed in this research was the argument of time savings for the respondent through the use of administrative records. In a follow-up paper to this research (Singer, VanHoewyk and Presser, 1999), the authors found "the quality of the data is a more important consideration than cost" and suggest that future research should "experiment with arguments that might be presented to the public in favor of data sharing."

An opportunity for this type of research presented itself in the spring of 2010, when the Census Bureau conducted a relatively small split-ballot field experiment on health insurance (called the Survey of Health Insurance and Program Participation, or SHIPP). The primary purpose of the SHIPP experiment was to examine three alternative sets of questions on health insurance coverage: one modeled on the Current Population Survey (CPS), one modeled on the American Community Survey (ACS) and one developed as an experimental design (called "EXP" hereafter). The SHIPP study included a request for consent to data linkage so that a follow-on validation study using administrative records could be carried out. Because there appeared to be no ideal candidate in terms of question wording for this consent question, three different versions of a consent question were developed in consultation with the literature, survey methodologists and staff from the Census Bureau policy office. Each version employed a different rationale for the data linkage: improved accuracy, reduced costs, and reduced respondent burden (wording and details below). The primary goal of the analysis is to assess any differences in levels of consent across these three versions. The secondary goal is an analysis of the demographics of those granting versus those denying consent. A third goal is an analysis of the respondents' reporting on key data needed to actually conduct the data linkage, once consent was granted.

Thus levels of missing data on address and date of birth for all household members were examined. Finally, a summary analysis examined the extent to which those in-scope for linking were different from those out-of-scope for linking – either because the respondent did not consent to linking, or did not provide key data needed to link.

The remainder of this paper is organized as follows: Section 2 includes wording of the consent questions, an overview of the SHIPP data collection methods, and details of the demographics questions. Section 3 provides the results; Section 4 is a discussion and summary; and Section 5 offers some thoughts for future research.

METHODS

Overview

The SHIPP survey was carried out from March 22 through May 10 of 2010 by the Census Bureau’s telephone interviewing staff in Hagerstown, Md., via three discreet but consecutive 10-day field periods. The survey was administered over the telephone using a CATI instrument and took an average of 17 minutes per household to complete (see Appendix A for details on the methodology). The content of the survey included basic demographics of all household members, disability, labor force participation and earnings, participation in government programs (such as Food Stamps), health insurance, a respondent debriefing, and finally the request for consent to link data. The sample was drawn from two sources – a Random Digit Dial (RDD) frame and Medicare enrollment files. The original sample size was 36,169, where 24,231 cases (67 percent) were RDD cases and the other 33 percent were Medicare cases. The goal was to complete 3,000 household interviews from the RDD sample and 2,000 interviews from the Medicare sample. That goal was exceeded for both sample types: there were 3,081 completed interviews from the RDD sample and 2,295 from the Medicare sample. In total these 5,376 households represented 12,743 people. Preliminary response rates (based on the AAPOR RR4 definition) were 47.6 percent for the RDD sample and 61.4 percent for the Medicare sample.

Advance Letter, Implicit and Explicit Consent

The SHIPP study was designed to mimic the conditions of the Current Population Survey (CPS) as much as possible. Regarding consent, in the CPS an advance letter is mailed which includes an explanation of linking plans, and it instructs respondents to inform the field representative during the interview if they do not want their data to be linked. If respondents do not object to linking during the interview, this is interpreted as “implicit consent” (sometimes called “informed consent” in related literature).

The SHIPP survey implemented this same general procedure for implicit consent where addresses were available, which was 100 percent of the Medicare sample but only 35 percent of the RDD sample. Across both types of sample, address was available for 56 percent of the

overall sample. These households were mailed a letter, and during the survey introduction respondents were asked if they received the letter. If they said yes, and did not inform the field representative that they objected to linking during the interview, this was considered implicit consent for the SHIPP study. If they did not say yes, an explicit request for consent was asked. Thus there were two conditions under which a respondent was asked for explicit consent: either they were mailed a letter but did not acknowledge receiving it, or they were not mailed a letter at all. When respondents were asked for explicit consent, one of these three versions of the consent question was asked, depending on which of the three field periods the respondent was interviewed (explained in more detail under “Experimental Design” below):

Field Period 1 (accuracy version): We'd like to conduct additional research to improve on the accuracy of our results by combining your survey answers with data from other government agencies. Do you have any objections?

Field Period 2 (cost version): We'd like to produce additional statistical data, and to reduce costs we'd like to combine your survey answers with data from other government agencies. Do you have any objections?

Field Period 3 (time version): We'd like to produce additional statistical data, without taking up your time with more questions, by combining your survey answers with data from other government agencies. Do you have any objections?

Each version of the question included an identical help screen:

HELP SCREEN

What Other Government Agencies and Why?

Your consent allows the Census Bureau to combine your answers with data we obtain from other government agencies, such as Medicare and Medicaid enrollment records. This helps make sure the data are complete and accurate.

Confidentiality Protections

The Census Bureau serves as the leading source of quality data about the nation's people and economy. We honor privacy and protect confidentiality, and all our research is conducted for statistical purposes only.

The same confidentiality laws that protect your survey answers – Title 13, US Code, Section 9 – also protect any additional information we collect. Providing your consent is voluntary.

Experimental Design

The study employed an inter-penetrated design whereby interviewers were assigned to one of three groups (later referred to as Interviewer Groups 1, 2 and 3, or IG1, IG2 and IG3) roughly balanced by skill and experience, based on field supervisors' judgment. Each group was then rotated across all three versions of the questionnaire during three separate, consecutive field periods. That is, once assigned to their group, interviewers were trained and worked on only one version of the questionnaire at a time, and at the conclusion of one field period they were rotated on to another version of the questionnaire and commenced the next field period of data collection. Thus all interviewers worked on all three questionnaire versions over the course of the survey. Due to the complexities of the design, however, a slight compromise was made with regard to the consent question. A fully inter-penetrated design would mean that each version of the linking question would be asked in each field period across all three health insurance panels

and interviewer groups. Instead, one version of the linking question (using accuracy as the rationale) was asked across all interviewer groups and all versions of the health insurance panels throughout the entire first field period. During the second field period, the cost version of the linking question was asked exclusively, and during the third field period the time version was asked exclusively. This confounds time with question wording of the consent question. We could see no reason time in and of itself would threaten the analysis, since the three field periods were all within a month of each other. However, learning curve could be a factor, as interviewers acquired experience with addressing respondent concerns over the life of the field period. If this was indeed factor, the accuracy version would suffer the most from lack of experience, and the time version would reap the most benefit, and the cost version would be in-between.

Collecting Name, Date of Birth and Address

While the manipulation of the linking consent question had to do with gaining permission to link, linking to administrative records goes beyond consent. It also involves collecting the necessary data elements, and obtaining high enough data quality on those elements, to successfully match a survey record to its corresponding administrative record. Those key data elements are generally name, date of birth and address. Analysis of missing data is a latter part of this analysis, so we briefly describe the methods of data collection for these elements.

The demographics section of the SHIPP survey was modeled after typical large household surveys. First a roster of all household members was collected, and then basic demographic data were collected for each individual (e.g.: date of birth, sex, marital status). Specifically, the name of each household member was requested, and collected in three discrete data fields (first, middle and last name). Date of birth for each person was also collected in three discrete data fields (birth month, day and year). If a valid month and year were provided, the instrument calculated the age as of the day of the interview and a question was asked to verify that age (in all cases respondents said “yes”). With regard to address, the state of residence was asked toward the end of the demographics section. For cases where address was known from the sample frame, no questions were asked to verify it. For cases where address was unknown, at the very end of the survey (after the linking question was asked), address was requested and collected in three discrete fields, in this order: zip code, city and street address. Thus in considering the impact of the consent question on willingness to provide key data, it is important to bear in mind that name, date of birth and state of residence are all asked before the consent question, and address is asked after the consent question.

Demographics

The survey included a fairly standard battery of demographic questions about each household member at the beginning of the questionnaire. Some of these (sex, age, race and education), as well as household income, are used as analytic and control variables in the analysis below. The question wording used to capture these data can be found in the questionnaire (available on request from the author). Some of these variables were restructured for ease of analysis, taking

into account both substantively meaningful categorizations and frequencies of originally-reported data. The race question contained five categories and respondents were told they could choose more than one race. That question was preceded by a yes/no question on Hispanic origin. Race and Hispanic origin responses were combined into one variable with the following four categories: (1) white non-Hispanic; (2) black non-Hispanic; (3) Hispanic; and (4) other (including those who chose multiple race categories, and those with missing data on race and Hispanic origin). The education categories were also regrouped, and collapsed into six categories: (1) less than a high school education; (2) high school graduate; (3) some college but no degree; (4) associates degree; (5) bachelors degree; and (6) professional degree. Toward the end of these person-level questions there was also a question on overall total household income. Respondents were asked whether the total combined income of all household members was above or below a certain dollar amount threshold. The dollar amount was based on household size and the presence of children and was meant to approximate the poverty level.

Evaluation Methods

We used several methods of evaluation. This paper focuses on response distributions, but other methods put in place were:

1. Interviewer debriefings: these were conducted at the conclusion of each of the first two field periods, just prior to training on the new questionnaire version. At the end of the third and final field period a debriefing was conducted on the third field period questionnaire version and the survey as a whole.
2. Interviewer diaries: a hard-copy question-level diary was provided to interviewers at the beginning of each field period and they were encouraged to write down any notable observations during the course of interviewing.
3. Tape recordings: approximately 100 interviews of each of the three questionnaire versions were tape recorded for later behavior coding analysis.
4. Validation of reported health insurance coverage through administrative records: future plans include linking the SHIPP survey data to Medicare records and possibly other sources of administrative data.

RESULTS

Linking Consent

1. Participation rates (n=12,338 sample units)

Given the rapid shifts in the use of data and records in general, and fears of identity theft, there is some concern that respondents who are informed of a study's plans to link data ahead of time may self-select out of the survey. In order to explore this we examine participation rates based on whether the household was mailed an advance letter. For this analysis we drop the out-of-scope cases since they were not eligible for participation, and thus the letter would play no role. Of the 36,169 total original sample cases, 23,831 (or 66 percent) were out-of-scope, leaving

12,338 cases for this analysis.

Among the in-scope cases, about three-fourths were mailed a letter (Table 1). Households that were mailed a letter had notably higher completion rates than households that were not mailed a letter (48 percent versus only 29 percent) and refusal rates were slightly lower among households that were mailed a letter (33 percent versus 35 percent). Thus participation was almost 20 percentage points higher in households that were mailed a letter, and there was no increase in refusals associated with the letter. This is not unexpected, regardless of the content of the letter. There is little evidence that respondents pay careful attention to the content of advance letters, and any potential negative effects of the notice on linking within the letter was apparently offset by the advantages of the letter, in terms of increased contact with the household and enhanced credibility. The non-contact rate is also much lower in households that were mailed a letter than in households that were not mailed a letter.

2. Implicit consent and opting out (n=5,376 households)

Of the 5,376 household respondents who participated in the survey, 2,058 (or 38 percent) gave implicit consent to link and were not asked the explicit linking question. Among those who were asked the explicit linking question, the vast majority did not object, resulting in a combined implicit/explicit consent rate of 90 percent. Respondents raised objections during the interview in only 35 cases and 28 of these interviews were conducted by a single interviewer. Though there was extreme bias by interviewer, the opt-out was observed across all three versions of the health questions and all three versions of the linking consent question. This indicates that the mechanism for opting out worked properly in all versions of the instrument, but that interviewers (at least one) may have failed to grasp the concept of this type of opt-out during training. Given the negligible frequency of this type of opting out, and the high concentration of the occurrence in a single interviewer, further analysis on this particular aspect of the interview has not been conducted.

3. Explicit consent (n=3,318 households)

The universe for this analysis is the group of respondents who were asked the explicit consent question (3,318 households). Overall, 84 percent of these respondents raised no objections to link their data, 14 percent did have objections, 2 percent refused the request and 1 percent said “don’t know.” These responses were collapsed to create a dichotomous dependent variable to express response to the linking consent request: no objections versus objections, don’t know, refused, as well as opt-out responses combined. Correlations between this dependent variable and various survey characteristics (such as the version of the consent question, sample type, etc.) and respondent demographics were tested. Chi-square tests were used for all bivariate analyses and logistic regression models were used for all multi-variate analyses.

4. Survey characteristics

Somewhat contrary to expectations, there was little difference in objections based on the version of the consent question. The cost version resulted in the highest level of “no objections” (85.3 percent), followed by the time version (83.6 percent), followed by the accuracy version (83.0 percent), but none of these differences was statistically significant in a chi-square test (chi-square=2.2; df=2; p=0.33). There were also no statistically significant differences in response to the linking request by sample type or version of the health questions. And there was no evidence that having been mailed a letter and not acknowledging it, versus not having been mailed a letter at all, made a difference in objection rates. There was, however, evidence of interviewer effects. IG2 had the highest rate of “no objections” at 86.5 percent, compared to 83.4 percent and 81.8 percent for IG3 and IG1, respectively (Table 2). To examine the effects of all these factors together, a logistic regression model was run with objection rate as the dependent variable and the following independent variables: version of consent question, sample type, interviewer group, version of health questions, and the letter condition. Only interviewer group was significant (chi-square=8.74; df=2; $p < 0.01$). IG1 was 1.21 times more likely to have an objection compared to IG2, while the difference between IG3 and IG1 was not significant.

5. Demographics of household respondent

The levels of objection varied based on the demographics of the household respondent (Table 3), and when all five demographic variables (age, race, sex, education and household income) were entered into a multi-variate model predicting objection to consent, race was not significant, but the other four factors were (Table 4). Older respondents were more likely to object; each year increase in age is associated with a 2 percent increase in the predicted odds of objecting, with a substantial shift at around age 56. Table 3 shows that those in each of the ~10 year age brackets under 56 years old objected at roughly the same rate, averaging 10.5 percent overall. But those aged 56-64 objected at a rate of 13 percent, and those 65+ objected at the highest rate of any age group, at 19 percent. Regarding income, those above and below the poverty threshold objected in equal proportions (15 percent), which was somewhat surprising. But consistent with other literature (Bates and Pascale, 2005), those with missing data on the income threshold question objected at a much higher level – 40 percent – and providing income data was associated with a 30 percent decrease in the odds of objecting. With respect to education, those with a professional degree or some college were least likely to object (14 percent). Those with less than a high school education were most likely to object (23 percent), and were 1.6 times more likely to object than those with some college. And finally, males were 1.13 times more likely to object than females though the difference in overall levels was modest (17 versus 16 percent).

As a final step, all five survey characteristics and all five respondent demographic variables were entered into the same model predicting objections to consent. Results (in terms of statistically significant predictors and coefficients) were largely unchanged from results of the two models being run separately: older respondents, those not responding the income question, males, those with less than a high school education, and those in IG3 were all more likely to object.

Missing Data Levels

As noted above, the SHIPP survey collected detailed information on name and date of birth for all household members, as well as household address. The current analysis is limited to missing data levels on date of birth and address only, as a larger coding effort will be required to assess the quality of name fields. For example, there is no readily-available code for incomplete or poor quality data on a name field; interviewers often type in “Lady of the House,” “LOH,” “second son,” etc. Thus each name would need to be examined by hand (or by a very sophisticated algorithm) in order to make some judgment about its quality and usefulness for linking purposes.

Because we want to examine the viability of linking survey data when we have permission, this analysis is limited to household respondents who were asked the explicit linking consent question and raised no objections; that is, 84 percent of 3,318 households, or 2,785 households representing 6,878 people. Valid data for state was reported in all cases. Overall 89 percent of the sample provided complete data for street address, city and zip code, another 6 percent provided only city and zip code, and 4 percent provided zip code only. Overall, complete date of birth (DOB) data were obtained for 83 percent of people in households that granted consent to link. For purposes of modeling a dependent variable, the address and DOB missing data variables were combined to create a single, dichotomous indicator of data quality: either all six data fields (three address and three DOB fields) had complete data, or they did not. Across all household respondents who were asked the linking consent question, 70 percent provided complete data and 30 percent did not.

1. Survey characteristics

The same variables on survey characteristics used in the objection analysis – version of consent question, sample type, interviewer group, version of health questions and the letter condition – were examined for any correlations with levels of missing data and results were similar to the above analysis. When all five variables were entered in a logistic regression model the only factor that was significant was interviewer group, and the same group (IG1) that had the highest objection rate also had the highest rate of missing data. IG2 was 1.12 times more likely to get complete data than IG1 (chi-square=4.5; df=1; $p < 0.03$). IG1 also had higher levels of complete data than IG3 and the difference approached statistical significance ($p < 0.07$).

2. Demographics of household respondent

Age, race, sex, education and household income were also examined for their associations with levels of missing data (Table 5). These variables were entered into a logistic regression model and results indicated that race, sex and education were unimportant, but age and non-response to the household income question were all associated with levels of missing data (Table 6). Similar to the objection findings, each year increase in age was associated with a one percent increase in the odds of providing incomplete data, and there was a marked jump in incomplete data among those in the oldest age category (65+). Those below 65 provided incomplete data 16-19 percent

of the time, but among those 65+ incomplete data was provided 25 percent of the time. For income, those above and below the income threshold provided incomplete data in roughly equal proportions (28 and 29 percent), but those with missing income data provided incomplete data 54 percent of the time, and providing income data was associated with about a 30 percent decrease in the odds of providing incomplete data on address and DOB.

And finally all ten variables were regressed on the missing data dependent variable and the outcomes were consistent with results from the two separate models in terms of the magnitude and direction of the significant coefficients. Older respondents, those not responding the income question, and those in IG3 were all more likely to provide incomplete data.

SUMMARY AND DISCUSSION

Summary

The key purpose of the experiment on different versions of the linking question was to examine whether different arguments or rationales for linking survey data with administrative records would resonate differently with respondents to the point that it affected their consent to data linkage. This experiment provides no evidence that question framing has an effect on consent. Consent rates varied from 83.01 percent to 85.27 percent, and none of the differences were statistically significant. Thus, somewhat contrary to our expectations, there is little evidence that respondents are influenced by the rationale given for linking – at least the rationales chosen in this study.

With regard to demographics and other survey characteristics, age of the household respondent and prevalence of income response were predictors of both response to the consent question and providing complete data on address and DOB. Older respondents (particularly those 65+) and those who failed to provide income data were most likely to object to the linking consent question and to provide incomplete data on address and DOB. Those with less than a high school education were also more likely to object to the consent question, and males were slightly more likely to object. There were also interviewer effects, where one of the three groups obtained higher objections rates and higher rates of missing data than their counterparts in another interviewer group.

Findings on demographics of the household respondent are, for the most part, consistent with related literature (Singer, Bates and Van Hoewyk, 2011), and expectations. Results suggest that the pool of people out of scope for linking (either due to objections by the household respondent and/or a failure to collect enough data to enable linking) are different from the pool of people in-scope for linking. Specifically, older respondents, those with less than a high school education and to some extent males would be out of scope for linking in higher proportions than their counterparts.

Context

The SHIPP consent experiment was carried out, in part, as an extension of a previous study using a similar research vehicle at the Census Bureau – the 2004 Questionnaire Design Experimental Research Survey (Bates and Pascale, 2005). In that study, consent to link was requested using wording very similar to the “time” version on the SHIPP study:

QDERS 2004: The Census Bureau would like to conduct additional research without taking up your time with more questions. We would like your permission to obtain the information that you have given to other government agencies on topics such as Social Security and Medicare benefits. Do you have any objections?

SHIPP 2010 (time version): We'd like to produce additional statistical data, without taking up your time with more questions, by combining your survey answers with data from other government agencies. Do you have any objections?

There are two key differences in question wording. First, QDERS referenced particular types of records (Social Security and Medicare) and SHIPP referenced the more general “data from other government agencies.” Second, QDERS asked for permission to *obtain* information from other agencies, while SHIPP asked to *combine* survey data with data from other agencies. The first change was a deliberate attempt to make the request more general and thus applicable to surveys on any topic. The second change was in the spirit of recent developments in the field that push for more “clear language” when requesting consent.

In the QDERS study, 63 percent of respondents had no objection – substantially lower than the 84 percent raising no objections in the SHIPP study – representing a difference of 21 percentage points over the six year time span. Study conditions in QDERS and SHIPP were similar; both were relatively small-scale, short CATI surveys on health insurance conducted out of the Hagerstown Telephone Center. Thus at least some factors driving the difference in levels of consent can be eliminated, and the more likely candidates are: (1) question wording; (2) survey climate (SHIPP was carried out at the peak of 2010 decennial field operations); (3) population (all respondents in QDERS were asked the consent question, but in the SHIPP study only those not in the implicit consent pool were asked the explicit question); (4) survey context and (5) time. Each will be discussed in turn.

It is very difficult to gauge the extent to which these wording differences played a role in the observed difference in consent levels. For the QDERS study, cognitive testing on variations of a consent question was conducted and results indicated that respondents had a very hard time grasping the concept of what it means to give permission to link their data, how the process actually works, what records and agencies would be involved, and so on (Bates and Pascale, 2005). While these results were less than promising, there was little in the findings (or other related literature for that matter) to suggest alternative wording that would improve respondent comprehension of the request. Since while no question wording candidate emerged as superior in terms of comprehension, respondents did find the longer versions to be somewhat burdensome and contribute nothing to the meaning so the final QDERS version was chosen primarily for its brevity. Thus while we cannot eliminate question wording as a factor in the QDERS-SHIPP differential in consent levels, there is not strong evidence that respondents understand these

types of requests in general. The relatively subtle differences between the QDERS and SHIPP wording, then, is unlikely to make a palpable difference in response.

With regard to survey climate, SHIPP data collection occurred during the height of the 2010 decennial census advertising campaign and operations, continuing through “Census Day” (April 1) and into early May. This has led to some speculation that the decennial census climate could have contributed to higher consent rates compared to 2004. There is some evidence that respondents may have confused the SHIPP survey for the decennial census and that this was reflected in response rates. Across the three field periods (March 22 - April 6; April 9 - 23 and April 26 - May 10) overall response rates were 52.8 percent, 50.8 percent and 51.2 percent respectively. One interpretation is that respondents in the first field period participated in the SHIPP mistakenly thinking it was the decennial census, then in the second field period respondents who already completed the real decennial census declined the SHIPP thinking it was a repeat decennial request, and by the third field period SHIPP response rates resumed a normal (that is, non-decennial-influenced) level. So while an argument could be made that the decennial affected *response rates*, it is unclear how and why the decennial climate would affect *consent levels*. Overall consent to the linking request fluctuated across the three field periods (83.01 percent, 85.3 percent and 83.6 percent respectively), but these differences were not statistically significant.

As for population, given differences in the overall approach to gaining consent in the two studies, there was a difference in the demographic profile of respondents who were asked the explicit consent question (Table 7). In QDERS, the request was asked of all respondents, but in the SHIPP only respondents who were not “skimmed off” with the implicit consent procedures were routed to the explicit consent request. That pool of respondents was somewhat different than the overall sample. Only households where an address was available were mailed a letter, and all cases in the Medicare sample contained an address, and the Medicare sample tended to be older than the overall sample. Consequently, respondents eligible for the implicit consent procedures were older than the overall sample, leaving a younger (presumably more cooperative) pool to be asked the explicit consent question. Indeed, in a logistic model containing all five demographic variables shown in Table 7, where the dependent variable was “version of consent question” (implicit or explicit), respondents in the explicit consent question pool were very slightly more likely to be younger (odds=1.009; chi-sq=16.9; df=1; p<0.001). One could argue that this difference in population composition could account for the QDERS-SHIPP gap in consent rates. However, in SHIPP the objection rate among the oldest, most opposed age group (65+) was still relatively low, with 82 percent responding “no objections.” So even a very crude adjustment for this bias – assuming the entire SHIPP explicit consent pool was made up of those 65+ – would result in an overall consent rate of 82 percent, which still results in a 19 percentage point gap in the QDERS 2004 consent rate of 63 percent.

Regarding survey context, while both the QDERS and SHIPP surveys asked about health insurance as their primary topic, SHIPP included questions on disability, labor force and program participation, while QDERS did not. Also, the set of questions immediately preceding

the consent question was different on the two surveys. In QDERS there was a set of questions on the details of any plans reported (such as the plan name, when it started, and who pays the premium). The SHIPP survey included a set of debriefing questions on respondents' confidence levels in the accuracy of their answers on the details of reported plans (such as the time period of coverage and the other household members covered). Interviewers reported in debriefings that respondents found these debriefing questions confusing and somewhat annoying. It is very difficult to speculate on how the respondent's state of mind would affect their comprehension of and answer to the consent question. On the one hand, respondents could have found it too taxing to focus on the meaning of the question and since the final phrase was "Do you have any objections?" a kind of reverse-acquiescence-bias could have set in and they just said "no." On the other hand, respondents truly annoyed with the final set of questions may have been disinclined to be cooperative and thus object to the consent question.

Finally, this brings us to the passage of time between 2004 and 2010. On first blush the notion that respondents are getting more cooperative over time regarding data linkage is inconsistent with other more general literature showing that respondents' trust in government and its use of administrative records has declined over the past 15 years or so (Singer, Bates and Van Hoewyk, 2011). However, as Singer et al note "a variety of evidence suggests that what people *say* they would oppose in a hypothetical situation does not necessarily predict what they would actually do in a real situation." Neither QDERS nor SHIPP asked hypothetical questions; both asked directly if respondents had objections to "us" (the Census Bureau) linking the data they had just provided in the survey to administrative records. So while the attitudinal and behavior trends seem to be going in different directions, this may be an artifact of some inherent disconnect between the way respondents answer hypothetical versus behavioral questions. Left to consider what the observed behavioral trend between 2004 and 2010 means, while most of the factors cited above cannot be ruled out in any definitive way, there is no compelling evidence to suggest that any of them made an important contribution to the difference in observed consent rates over time. Thus a very tentative conclusion from this research is that there is evidence of a real shift in the public's willingness to allow the government to use administrative records.

Interviewer Effect

Lastly, considering the results on interviewer effects for both outcomes of interest (rates of objection to the consent request and rates of missing data on key linking variables) we examine the relationship between these outcomes and the two characteristics of SHIPP interviewers available on the dataset – the number of years interviewing in general, and on health surveys in particular (Table 8). IG2 achieved the most favorable results on both outcomes of interest and, perhaps contrary to expectations, this group was in the middle of the other two groups in terms of experience - 7.2 years compared to 5.6 for IG3 and 8.3 for IG1. Perhaps most surprising is that IG1, which had the lowest outcomes on both measures of interest, had the most experience in terms of years interviewing in general, and also on health surveys.

FUTURE RESEARCH

There are several avenues for further research on the SHIPP consent questions. Behavior coding of the tape recordings, and an examination of the interviewer diaries, could shed some light on respondent comprehension of and concerns about the linking request. Cognitive testing of the questions (and/or similar versions) could also be conducted to assess whether respondent comprehension has improved, and possibly to give some perspective to whether the observed difference between QDERS 2004 and SHIPP 2010 is due to wording differences, attitudinal changes over time, other factors, or some combination.

Given the relatively high overall consent rate of 90 percent, and the fact that complete address data were obtained for 89 percent of households, and complete DOB data were obtained for 83 percent of persons in interviewed households, further research on non-response bias in these particular data could be warranted. This could consist of several phases: (1) a coding operation to assess the quality of the data on names; (2) a statistical research operation to assess the overall match rate between the survey data and administrative records, given the data items available at the individual level; and (3) an assessment of non-response bias between matched and unmatched individuals. In more general terms, future research could involve the development of tailoring strategies – in terms of survey participation recruitment efforts and data collection methods – and post-survey adjustments for the sub-populations mostly likely to be ineligible for linking.

REFERENCES

- Bates, N., & Pascale, J. (2005). *Development and testing of informed consent questions to link survey data with administrative records* (Unpublished report). Washington, DC: U.S. Census Bureau, Demographic Surveys Division and Statistical Research Division.
- Singer, E., Bates, N. & Van Hoewyk, J. (2011, May). *Concerns about privacy, trust in government, and willingness to use administrative records to improve the decennial census*. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix Arizona.
- Singer, E. & Presser, S. (1996). *Public attitudes toward data sharing by federal agencies* (Report on the 1995 Joint Program in Survey Methodology Practicum). Washington, DC: U.S. Census Bureau.
- Singer, E., Van Hoewyk, J. & Presser, S. (1997). Proceedings from the Record Linkage Techniques – 1997 International Workshop and Exposition: *Public attitudes toward data sharing by federal agencies*. Arlington, VA: National Academy Press.

Table 1: Sample Disposition by Mailed Letter Status (In-scope cases only; n=12,338 telephone numbers)

Sample Type	Completed Interview		Refusal		Non-Contact		Total	
	n	%	n	%	n	%	n	%
Letter Mailed	4,551	48.03%	3,082	32.53%	1,842	19.44%	9,475	100%
No Letter Mailed	825	28.82%	1,002	35.00%	1,036	36.19%	2,863	100%
Total	5,376	43.57%	4,084	33.10%	2,878	23.33%	12,338	100%

Table 2: Consent Question Response by Survey Design Characteristics (n=3,318 households)

			Explicit Request for Consent		Total
			No objections	Yes objects/DK/REF	
Overall		n	2,785	533	3,318
		%	83.94%	16.06%	100%
Linking Version	Accuracy	n	933	191	1,124
		%	83.01%	16.99%	100%
	Cost	n	903	156	1,059
		%	85.27%	14.73%	100%
	Time	n	949	186	1,135
		%	83.61%	16.39%	100%
Sample	RDD	n	1,702	331	2,033
		%	83.72%	16.28%	100%
	Medicare	n	1,083	202	1,285
		%	84.28%	15.72%	100%
Interviewer Group	Group 1	n	878	195	1,073
		%	81.83%	18.17%	100%
	Group 2	n	962	150	1,112
		%	86.51%	13.49%	100%
	Group 3	n	945	188	1,133
		%	83.41%	16.59%	100%
Health Q Version	CPS	n	965	164	1,129
		%	85.47%	14.53%	100%
	ACS	n	954	188	1,142
		%	83.54%	16.46%	100%
	EXP	n	866	181	1,047
		%	82.71%	17.29%	100%
Letter	Letter Mailed (did not receive)	n	2078	414	2,492
		%	83.39%	16.61%	100%
	No letter mailed	n	707	118	825
		%	85.70%	14.30%	100%

Table 3: Consent Question Response by HH Respondent Demographics (n=3,318 households)

			Explicit Request for Consent		Total	
			No objections	Yes objects/DK/REF		
Age	17-25	n	40	4	44	
		%	91%	9%	100%	
	26-35	n	145	18	163	
		%	89%	11%	100%	
	36-45	n	307	34	341	
		%	90%	10%	100%	
	46-55	n	474	57	531	
		%	89%	11%	100%	
	56-65	n	627	95	722	
		%	87%	13%	100%	
	65+	n	934	214	1,148	
		%	82%	19%	100%	
	Household Income	Above poverty	n	1,800	309	2,109
			%	85%	15%	100%
Below poverty		n	882	156	1,038	
		%	85%	15%	100%	
Missing		n	103	68	171	
		%	60%	40%	100%	
Education	Professional degree	n	332	53	385	
		%	86%	14%	100%	
	Bachelor's degree	n	522	90	612	
		%	85%	15%	100%	
	Associates degree	n	236	48	284	
		%	83%	17%	100%	
	Some college, no degree	n	533	90	623	
		%	86%	14%	100%	
	High school grad	n	922	174	1,096	
		%	84%	16%	100%	
	Less than high school	n	214	65	279	
		%	77%	23%	100%	
Sex	Male	n	1,038	212	1,250	
		%	83%	17%	100%	
	Female	n	1,745	321	2,066	
		%	84%	16%	100%	
Race	White non-Hispanic	n	2,226	424	2,650	
		%	84%	16%	100%	
	Black non-Hispanic	n	277	57	334	
		%	83%	17%	100%	
	Hispanic	n	125	18	143	
		%	87%	13%	100%	
	Other	n	157	34	191	
		%	82%	18%	100%	

Table 4: Consent Question Response by HH Respondent Demographics: Logistic Model Results (modeling the probability that the respondent objected)

Variable		Estimate	SE	Odds Ratio
Age		0.018*	0.004	1.02
HH income (omitted=non-response)	Above	- 0.36*	0.090	0.70
	Below	- 0.38*	0.097	0.69
Education (omitted=some college)	Professional degree	- 0.21	0.153	0.81
	Bachelor's degree	- 0.21	0.130	0.81
	Associate's degree	0.16	0.158	1.18
	High school graduate	- 0.05	0.097	0.95
	Less than high school	0.46*	0.148	1.59
Female		- 0.12*	0.056	0.89
Race (omitted=White non-Hispanic)	Black non-Hispanic	0.12	0.162	1.13
	Hispanic	-0.24	0.245	0.79
	Other	0.05	0.204	1.05
Intercept		- 2.58	0.265	0.08

* p < .05

Model chi-sq likelihood ratio=80.09; df=12; p < .0001

Table 5: Missing Data Levels by HH Respondent Demographics (n=3,318 households)

			Levels of DOB and Address Missing Data			
			All six fields good	At least 1 field missing	Total	
Age	17-25	n	37	7	44	
		%	84%	16%	100%	
	26-35	n	136	27	163	
		%	83%	17%	100%	
	36-45	n	282	59	341	
		%	83%	17%	100%	
	46-55	n	432	99	531	
		%	81%	19%	100%	
	56-65	n	585	137	722	
		%	81%	19%	100%	
	65+	n	856	292	1,148	
		%	75%	25%	100%	
	Household Income	Above poverty	n	1,506	603	2,109
			%	71%	29%	100%
Below poverty		n	744	294	1,038	
		%	72%	28%	100%	
Missing		n	78	93	171	
		%	46%	54%	100%	
Education	Professional degree	n	273	112	385	
		%	71%	29%	100%	
	Bachelor's degree	n	411	201	612	
		%	67%	33%	100%	
	Associates degree	n	200	84	284	
		%	70%	30%	100%	
	Some college, no degree	n	444	179	623	
		%	71%	29%	100%	
	High school grad	n	798	298	1,096	
		%	73%	27%	100%	
	Less than high school	n	191	88	279	
		%	68%	32%	100%	
Sex	Male	n	860	390	1,250	
		%	69%	31%	100%	
	Female	n	1,467	599	2,066	
		%	71%	29%	100%	
Race	White non-Hispanic	n	1,857	793	2,650	
		%	70%	30%	100%	
	Black non-Hispanic	n	240	94	334	
		%	72%	28%	100%	
	Hispanic	n	104	39	143	
		%	73%	27%	100%	
	Other	n	127	64	191	
		%	66%	34%	100%	

Table 6: Levels of Missing Data by HH Respondent Demographics: Logistic Model Results (modeling the probability that at least one data field is missing)

Variable		Estimate	SE	Odds Ratio
Age		0.01*	0.003	1.01
HH income (omitted=non-response)	Above	- 0.30*	0.080	0.74
	Below	- 0.40*	0.088	0.67
Education (omitted=some college)	Professional degree	- 0.13	0.127	0.88
	Bachelor's degree	0.04	0.104	1.04
	Associate's degree	0.01	0.141	1.01
	High school graduate	- 0.16	0.085	0.85
	Less than high school	0.25	0.137	1.28
Female		-0.07	0.048	0.934
Race (omitted=White non-Hispanic)	Black non-Hispanic	-0.08	0.144	0.924
	Hispanic	-0.03	0.195	0.967
	Other	-0.01	0.175	0.99
Intercept		01.93	0.222	0.15

* p < .05

Model chi-sq likelihood ratio=65.15; df=12; p < .001

Table 7: Demographic Characteristics of All Household Respondents (n=5,376 households)

			Gave Implicit Consent	Asked Explicit Consent Question	
Age	17-25	n	19	44	
		%	1%	1%	
	26-35	n	84	163	
		%	4%	5%	
	36-45	n	144	341	
		%	7%	10%	
	46-55	n	328	531	
		%	16%	16%	
	56-64	n	464	722	
		%	23%	22%	
	65+	n	897	1,148	
		%	44%	35%	
	Household Income	Above poverty	n	1,461	2,109
			%	71%	64%
Below poverty		n	509	1,038	
		%	25%	31%	
Missing		n	88	171	
		%	4%	5%	
Education	Professional degree	n	301	385	
		%	15%	12%	
	Bachelor's degree	n	395	612	
		%	19%	18%	
	Associates degree	n	149	284	
		%	7%	9%	
	Some college, no degree	n	390	623	
		%	19%	19%	
	High school graduate	n	666	1,096	
		%	32%	33%	
	Less than high school	n	140	279	
		%	7%	8%	
	Sex	Male	n	763	1,250
			%	37%	38%
Female		n	1,295	2,066	
		%	63%	62%	
Race	White non-Hispanic	n	1,814	2,650	
		%	88%	80%	
	Black non-Hispanic	n	118	334	
		%	6%	10%	
	Hispanic	n	56	143	
		%	3%	4%	
	Other	n	70	191	
		%	3%	6%	
TOTAL		n	2,058	3,318	
		%	100%	100%	

Table 8: Interviewer Characteristics, Consent Rates and Missing Data Rates

	Interviewer Experience		Survey Outcomes	
	General	Interviewing on	No objections	All six fields
Group 1	8.3	6.9	82%	65%
Group 2	7.2	4.2	87%	73%
Group 3	5.6	3.0	83%	72%
AVERAGE	7.0	4.7	84%	70%