Matching surveys in longitudinal studies

Cynthia B. Augustine[1], Scott Ginder[1], Kristine L. Rae Olmsted[1]
[1]RTI International 3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract
This research addresses the issue of linking baseline and followup surveys in longitudinal studies. In the past, Social Security Number (SSN) was commonly used as a matching variable between survey waves, but reluctance to provide this information reduced matches. Optical scanning errors can also produce mismatches between otherwise matching surveys. Two matching methods were developed for a smoking survey that had no other identifying information except SSN. The variable matching methods were explored using a similar smoking survey on a parallel population. Some initial conclusions from this case study are presented to guide further research.

Key Words: Longitudinal, tobacco, alcohol

1. Background and Problem

In late 2005 and early 2006, a baseline survey of tobacco use was administered to a population located at four locations in the United States. This survey will be referred to as the Tobacco Survey. The baseline was completed by 5,534 participants. Six weeks later, a followup survey with nearly identical question items was administered to participants from the baseline survey who were still located at their original location. The followup survey had 2,262 respondents for a crude, non-matched response rate of 40.87%. This low response rate was further complicated when a number of baseline and followup surveys were unable to be matched. A self-reported social security number (SSN) was the sole variable linking the survey waves; some respondents reported inaccurate SSNs or did not report an SSN at all. Further, the optical scanning equipment and software would somewhat frequently misinterpret illegible handwriting such that true matches were unable to be directly linked to one another.

Two main methods of linking the unmatched surveys were implemented. The first method involved matching survey waves on birthdate; marital status; and response to the questions “Have you ever smoked a cigarette, even a puff,” “Have you ever smoked everyday for a month,” and “Have you smoked in the last month?” If responses to two out of four items, in addition to birthdate, matched across surveys, the survey ID numbers were noted and an analyst reviewed an image file of the survey to see if illegible handwriting and/or poor optical scanning quality resulted in the discrepancy. The second method examined the SSNs for strings of matching digits. For instance, baseline and followup surveys with a string of seven digits in common and in order, would be visually examined to see if the optical scanner could have misinterpreted the person’s handwriting on the remaining two digits, thus invalidating a match. Using these methods, it was possible to raise the response rate from an initial matching response rate of 16.47% to 22.70%.

These methods may be useful for other surveys where linking survey waves is difficult due to scanning errors, so as to increase response rates. The variable linking method was evaluated on a similar population across three locations during a similar timeframe. The second survey in this example will be referred to as the Alcohol and Tobacco Survey. The Alcohol and Tobacco Survey contained many of the same question items as the Tobacco Survey, but included additional items on alcohol use. Legitimately matched waves in the Alcohol and Tobacco survey were examined for variable response differences to assess the effectiveness of the variable matching method,

2. Tobacco study

2.1 Variable Matching
We implemented a matching scheme based on variables for birthdate, marital status, “Have you ever smoked a cigarette, even a puff”, “Have you ever smoked everyday for a month” and “Have you smoked in the last month”. There were several iterations of matching where the team explored other possible variables for potential use in

matching baseline to followup surveys. The variables for mother's education, father's education and self-education were also explored to help identify matches, but these were not reliable. The variable match process identified 222 matches of the baseline and followup surveys.

2.2 Social Security Number Matching

We wrote a SAS program to identify SSNs that were of specified numbers of digits different from the baseline to the followup; matching on numeric strings of 5, 6, 7, and 8 digits was used to identify potential matches. It was determined that strings of 2, 3 and 4 numbers were insufficient to produce accurate results. Strings of matching SSNs were visually verified (using image files of the original surveys). Pairs of SSNs with fewer matching digits were visually inspected more often than SSN strings with 7 or 8 matching digits. Strings of 5 or 6 matching digits were more likely to have several possible matches and required more visual inspection to find the truly matching SSN. *Table 1* below contains some examples of the SSN strings (Note: Decimals indicate the digits were the same).

Table 1*:* Examples of Matching SSN Strings

Baseline	**Followup**
90.....	96.....
.....483	983
.....869	569
.....419	919
.....959	459
.....062	862
.....875	375

The sole use of handwriting (i.e., not fill-in bubbles) for the social security number on this study increased the likelihood of errors in this matching variable. *Table 1* clearly illustrates that zeroes were misinterpreted as sixes, fours as nines and eights as fives, as well as errors with other digits.

3. Alcohol and Tobacco study

3.1 Similarities and Differences

The tobacco questions in the two surveys were nearly identical; the Alcohol and Tobacco study contained additional questions relating to alcohol use. Most of the same demographic information from the Tobacco study was collected for the Alcohol and Tobacco study. The two surveys were nearly the same length, with the Alcohol and Tobacco study being slightly longer in duration due to the addition of alcohol items. The Alcohol and Tobacco Study matched survey waves using a written and bubbled SSN, whereas the Tobacco study relied solely on the subject's handwriting for a written SSN. The Tobacco study examined a baseline and a 6-week followup where the Alcohol and Tobacco study examined a baseline and a 4-month followup. The baseline surveys for both studies were paper and pencil and were administered in a group setting at each location; locations for the two studies differed. The followup surveys differed in administration method: the Alcohol and Tobacco Survey was web-based whereas the Tobacco Survey remained a paper and pencil survey administered in a group setting.

The Alcohol and Tobacco study had 6,298 baseline respondents. Only 5,061 respondents remained when duplicate and missing SSNs were removed. The web-based followup survey had 1,810 completed responses.

3.2 Variable Matching

For the Alcohol and Tobacco study, 1,656 baseline and followup surveys matched automatically on SSN alone. Using those matched surveys, we implemented variable matching scheme to investigate the level of missingness of key variables used as part of this method. By focusing solely on baseline and followup surveys, we were able to evaluate the effectiveness of the method in identifying potential matches.

4. Results

4.1 Variable matching process

Our examination of items asked in both baseline and follow-up surveys for the legitimately matched cases in the Alcohol and Tobacco survey identified areas of potential concern with the variable matching method used in the Tobacco survey. For example, there were high levels of item nonresponse in the followup survey for the Alcohol and Tobacco survey. Additionally many of the responses to these key questions were different at the baseline and the followup. Missing values in the key variable were excluded when looking for changes between baseline and followup. For example, marital status changed for 37% of the respondents. The question regarding regular smoking in the past month was different for 88% of the respondents and the question regarding smoking in the past month was different in 66% of the instances. The high missingness and the change in responses indicate that the variable matching process might not have been reliable in this example. One reason for the change in responses may have been the longer time between the baseline and the followup. Four months had elapsed, and a different survey mode was used, which may have affected the response outcomes. The matching process in the Tobacco study may have been more reliable due to the shorter followup time.

4.2 Subgroup Results for the Tobacco study

In the Tobacco Survey the matching process yielded slightly more males than females. This was not unexpected since the majority of this population was male, but it could indicate that females are less likely to provide SSN. There were no major changes in proportions of ethnicity, marital status or education. The matched surveys saw small increases across the racial groups. The more prevalent races (e.g. white and black) had larger increases than other smaller racial groups. The Asian-American Indian and Hawaiian-Pacific groups increased. Due to the racial makeup of the population, it was not expected that these would increase substantially. The smoking variable responses were nearly identical to the pre-match proportions. The variable matching process improved overall precision by increasing the number of matches.

5. Discussion

5.1 Reluctance to provide SSN

Given stories in major media outlets over the past few years about identity theft, the American public is reluctant to provide an SSN. The blank and missing SSNs on the Tobacco survey were evenly distributed amongst variables of interest. This included demographic and smoking characteristics. This suggests that reluctance to provide SSN was not related to demographic characteristics nor was it related to tobacco use.

5.2 Other considerations

One of the considerations prompting this exploratory case study of two matching processes was the concern that collecting data from participants and then not using may be considered unethical. Both surveys were somewhat lengthy and the participants were not compensated for their time. However, in the case of these surveys it does not seem appropriate to spend time testing and developing a matching process for such little improvement in precision, when the overall response rate is already so low.

5.3 Recommendations

It is advisable to avoid using SSN to match survey waves if at all possible; recent regulations regarding use of protected health information (PHI) make this point. Social Security Number matching and variable matching improved overall response in the Tobacco Survey, but the effectiveness of the variable matching process is not confirmed due to high levels of missingness.

For variables that will be used to identify matching baseline and followup surveys, bubbled response options should be used whenever possible, as opposed to hand-written responses. Researchers could consider using a respondent created ID variable (Dillman, 2000) or a randomly-generated identification code that is the same for a baseline survey and a followup.

5.4 Conclusions

Future research should focus on methods to improve overall response rates at both baseline and followup. Further research should establish better techniques for matching waves of longitudinal surveys. Protections have been implemented to restrict use of SSN, thus waves will need other reliable linking variables. Surveys with higher response rates are more representative of the population and a more effective use of resources. Subgroup analyses maybe considered with higher response rates.

References

Dillman, Don A. 2000. “Mail and Internet Surveys: The Tailored Design Method.” New York: Wiley.