Updated 8/15/2002

VLS PROJECT TECHNICAL MEMO # 2

Tracking and Matching Households and Persons Across Rounds of the VLS

OVERVIEW

The VLS incorporates a longitudinal research design with the baseline 1995 sample of households and persons contacted and reinterviewed in 1996, 1997, and 1998. In the VLS, as in all longitudinal studies, there are two problems that threaten the completeness and quality of the longitudinal data. The first is the inevitable attrition of respondents from the baseline survey as households are disbanded, respondents die or move to different places. The second problem is that some erroneous matches might be made across rounds of the survey, that is, different persons are assigned the same unique ID, which is used to match persons in the merged data file. In this memo, we explain the potential sources of these errors and the procedures used in the field and in the office to discover and minimize errors. We also present information on the "corrected" merged data file, which was produced after the evaluation of the completeness and quality of the matched records of households and persons across survey rounds.

In the follow-up rounds (1996, 1997, and 1998) of the VLS, interviewers were trained to locate the households and individuals first contacted and interviewed in the baseline 1995 VLS. Interviewers had household roster information from the 1995 survey in order to insure the same persons were reinterviewed. In rural areas, there was relatively little turnover in households, except for occasional deaths, births, and the addition of new spouses. The major reason for the loss of respondents was outmigration. Although efforts were made in the later rounds of the VLS to track persons who migrated, many out-migrants were not contacted and reinterviewed. One technique used to minimize sample attrition was to obtain information about missing respondents from other persons in their households. These "proxy interviews" were a last resort when the missing respondent could not be contacted and reinterviewed directly.

In the VLS, each of the ten rural and urban areas (communes in rural areas and subareas in towns) had a unique ID number. Then, each household in an area was given unique (consecutive) household ID number. Within each household, each person had a unique person (line) number. In theory, each household and person can be matched with these unique numbers. Because of changes in household composition, it is possible that the listing of persons (and associated person numbers) changed across rounds of the survey. It is also possible that the interviewers made errors in entering responses during the interview (or coding). These actions may lead to errors where the same ID number is used for different persons across survey rounds or cases where the person is the same, but their characteristics are reported inconsistently across rounds.

In our preliminary evaluation of the VLS data, we discovered that some "matched persons" reported different years of birth and even a few cases with different gender for the same person across survey rounds. These differences could arise because of incorrect matches of persons, or because of response errors in one or more surveys. For example, a respondent may have reported an incorrect date of birth or a coder could have misread a response in data entry. To evaluate the magnitude of sample attrition and the quality of tracking respondents, we carefully examined the characteristics of matched records across survey rounds. Our objectives are (1) to verify whether the same households and persons are matched across the four waves of interviews, and (2) to correct reporting errors in sex and year of birth of interviewed persons after ensuring that these cases were matched correctly.

DATA EVALUATION AND ADJUSTMENT

The first step of the evaluation was to match each household and person across each survey and to examine the commune ID, the household ID, the line number of household informants in household file, and the line number of respondents in individual file. With these data, we compared the names and basic demographic characteristics of households and persons for each survey round. For each household that has a discrepancy in the age or sex of a respondent, the entire list of household members, including names, sex, and year of birth were examined. Based on this detailed examination, it was possible to determine two types of errors: The first type was a mistaken ID number (ICASENUM), when the matched record (same ICASENUM) was not the same person. Usually, the correct person for the match could be identified from the listing of all persons in the household. The second type of error was when the same person was matched, but their personal characteristics were coded differently (e.g. year of birth) across survey rounds.

The first type of error, different persons sharing the same ID number, were identified only after a detailed comparison of the following variables: commune ID, household ID, line number in household roster, name, year of birth, sex, relationship with the head of household, residential status, and marriage status. Only after conclusively identifying an incorrect match, were corrections made to adjust ICASENUM, to that the same persons are matched across years.

Correcting the second type of error is more difficult because some inference must be made about reporting errors from the interview (or coding). For example, it appears that there were two cases where the correct person was identified but the respondent's gender differed across survey rounds. We were able to identify their correct gender based on their Vietnamese name. For discrepancies of age, we had to look for cues from other variables including names, years of birth of their siblings, years of marriage and age of the spouses at that time, and years of birth of parents and children. If these items did not allow an unambiguous correction to be made, we relied on the following assumptions to correct reporting errors: the 1995 baseline responses were considered more reliable than responses from later rounds and; data from individual (direct) interviews were considered more reliable that data from household records (based on indirect reports)

QUALITY OF THE MERGED FILE

Based on the methods described above, approximately 440 persons in about 320 households, which have been adjusted (corrected). Almost all of these corrections were in the year of birth of respondents. Only two household numbers and 22 households with incorrect person line numbers were found to be in error and corrected. In two cases, we concluded that the gender of the respondent was wrongly coded. Most of the errors in measuring birth were fairly small (1-3 years different) and were most common in 1997 and 1998.

The following two tables that show the numbers of households and persons that were successfully contacted and re-interviewed across the four rounds of the VLS. Following the tables is an appendix with syntax (computer program) used to evaluation and correct the VLS data.