NLSY79 APPENDIX 10:

1979–2000 GEOCODE DOCUMENTATION

Documentation for the NLSY79 Geocode Data Files

This document provides a discussion on the creation of the variables available on the NLSY79 geocode data files. These supplemental data files provide selected variables from the *County and City Data Books* from various years along with geographic variables from the NLSY79 main data file.

The county and state of residence for each NLSY79 respondent for each survey year were matched with the county and state variables on the *County and City Data Book* data files used for those years, and selected county-level or SMSA-level environmental variables were extracted from those files and included on the geocode data files.

> ***User Notes:*** Two versions of the county and state of residence variables are included in the Geocode data files for most survey years from 1979–92. The state and county variables appearing at the beginning of each file are the edited versions that incorporate all revisions deemed necessary in the hand-editing process for each year. These edited variables are used in the construction of the final geocode data files. The state and county variables appearing near the end of the file for most years are the unedited version, as received directly from NORC. It is generally recommended that users employ the edited version as these contain corrected geocodes based upon the most current available information.

The *County and City Data Book* data files were prepared by the U.S. Census Bureau. Related printed matter for each of these data files can be found in the *County and City Data Book* for the specified year. These books are also published by the Census Bureau.

The following is a brief description of the various NLSY79 geocode data files and the *County and City Data Book* data files that were merged with the different years of NLSY79 data:

1. The 1979–82 geocode data files include county-level and SMSA-level variables from the *County and City Data Book*, 1972 data file, which provides data from the *1970 Census of the Population and Housing*, the *1972 Economic Census*, and the *1969 Census of Agriculture*, and other data derived from a variety of federal government and private agencies.

2. The 1979–82 geocode data files include county-level and SMSA-level variables from the *County and City Data Book*, 1977 data file, which provides data from the *1970 Census of the Population and Housing*, the *1972 Economic Census*, and the *1974 Census of Agriculture*, and other data derived from a variety of federal government and private agencies.

3. The 1983–87 geocode data files include county-level variables from the *County and City Data Book*, 1983 data file, which provides data from the *1980 Census of the Population and Housing*, the *1977 Economic Census*, the *1978 Census of Agriculture*, and other data derived from a variety of federal government and private agencies.

4. The 1988–96 geocode data file includes county-level variables from the *County and City Data Book*, 1988 data file which provides data from the *1980 Census of the Population and Housing*, the *Current Population Surveys*, and other data derived from a variety of federal government and private agencies.

5. The 1998–2000 geocode data file includes county-level variables from the *County and City Data Book*, 1994 data file, which provides data from the *1990 Census of the Population and Housing*, the *Current Population Surveys*, and other data derived from a variety of federal government and private agencies.

Another type of data file, the City Reference File (CRF) for various years, was also merged with the NLSY79 data in order to identify the SMSA/MSA for each respondent according to zip code. The City Reference File data files, prepared by the U.S. Census Bureau, contain the Federal Information Process Standards (FIPS) county and state codes, zip codes, and SMSA/MSA codes.

The following is a list of the various City Reference Files that were merged with the different years of NLSY79 data to identify the SMSA/MSA for each respondent:

1. The 1979–82 NLSY79 data was merged with the City Reference File, 1973, which contains the SMSA codes as defined by the Office of Management and Budget (OMB) as of August 15, 1973.

2. The 1983 NLSY79 data was merged with the City Reference File, 1982, which contains the SMSA codes defined by OMB prior to June 30, 1983.

3. The 1984–87 NLSY79 data was merged with the City Reference File, 1983, which contains the MSA codes as defined by OMB as of June 30, 1983.

4. The 1988–92 NLSY79 data was merged with the City Reference File, 1987, which contains the MSA codes as defined by OMB as of June 30, 1987.

5. The 1993–2000 NLSY79 data was merged with the City Reference File, 1993, which contains the MSA codes as defined by OMB as of July 31, 1993.

Beginning in 1989, a third type of data file was used to verify the geocode information provided by NORC for each respondent. The Local Exchange Routing Guide (LERG) data file is constructed by Bell Communications Research (BELLCORE) and contains address information for the "switches" which regulate each telephone area code and exchange. The LERG data file used for the 1989 NLSY79 Geocodes was updated through October 1989. The LERG file used for the 1990 and 1991 NLSY79 Geocodes was updated through January 20, 1992. The LERG file used for the 1992 NLSY79 Geocodes was updated through March 1, 1993. The LERG file used for the 1993–94 NLSY79 Geocodes was updated through August 1, 1994.

In order to facilitate merging of the NLSY79 geocode data files with the NLSY79 main data files, the records on all files are sequentially ordered according to the respondent's identification number (ID). Additionally, the respondent's ID is included on each NLSY79 geocode file and can be used to create an extract file containing only a subset of sample cases. The number of records on each of the files is 12,686.

1979–82 Geocode Data File Creation Procedures

The following briefly outlines the procedures used to create the initial 1979–82 NLSY79 geocode files.

1. State, county, and zip codes are reported by each NLSY79 respondent. Missing information was hand-edited whenever possible. (See the discussion below on hand-

edits.) A variable was created to indicate the type of hand-editing that was done on each case.

2. The state, county, and zip codes were then matched with the 1973 CRF. For those cases where the NLSY79 state, county, and zip codes matched with a state, county, and zip code from the CRF, the SMSA from the CRF was added to each respondent's record.

3. The NLSY79 file, with SMSA added when there was a match on all three residence variables, was then merged with the *County and City Data Book*, 1972 and with the *County and City Data Book*, 1977 data files.

Hand-Edits

More than 1,000 hand-edits for each survey year from 1979 through 1982 were performed to constrain respondents' reported state, county, and zip codes so that they conformed to legitimate state-county-zip combinations. In some cases, this involved making estimates for one or more of the above items. The state and county codes from the main NLSY79 data for each year that are included on the NLSY79 geocode files are the original, unedited values for the respondents. The hand-edited versions of the state and county codes were used to match with the CRF and with the *City and County Data Book* data files for these years.

In compiling the 1981 geocode information, a systematic review of hand-edited state, county, and zip codes was also undertaken. All cases that required a hand-edit in any of the three survey years were included in this inspection. The point of this review was: (1) to check for consistency in hand-editing decision rules over the three years, and (2) where possible, to use the respondent's reported geocodes in subsequent years to check on the accuracy of hand-edits performed in preceding years (this was possible for those cases that required hand-edits in early years and which showed no change of residence over the period).

The results of these consistency checks were very encouraging. Only 13 cases turned up that seemed to be in error. These cases had their geocodes revised accordingly. While doing this review, several dozen other cases with keypunch or coding errors *in the hand-edit code variables* were also uncovered. These errors were also corrected. In any case, this procedure provides substantial validation to the overall hand-editing process.

1981 Changes in SMSA Designations

For those using these data to track the mobility of respondents over the 1979–82 survey years an additional caution applies. In June of 1981, the OMB announced the designation of 36 new SMSAs, the disqualification of one pre-existing SMSA, and the merger of two pre-existing SMSAs into one new area. The 1973 CRF file was updated by CHRR to reflect these changes, and the updates were applied beginning with the 1980 interview place of residence in the 1980 geocode data file. One consequence of these changes is that when attempting to match places of residence for respondents using data from a 1979 geocode file and a separate 1980 update geocode file, some respondents give the appearance of moving into (or out of) an SMSA between 1979 and 1980 when in fact they may not have moved at all. This faulty inference of mobility would be reached if one compared changes in SMSA designation between the separate 1979 and 1980 update geocode data files.

Users ordering a full compliment of geocode files at any given point should not find this discrepancy in mobility. This applies only to those who ordered a 1979 geocode file and then updated that data with single year files in subsequent years. As single year files were no longer available after the 1979–89 release, recent purchasers of the geocode data would have received all

available years of the geocode data, and should not detect the discrepancy resulting from the 1981 SMSA changes between the 1979 and 1980 separate data files. A variable representing the 1981 SMSA designation (if applicable) of place of residence at interview is currently present in the geocode data for all survey years, including 1979.

It is possible however that the created variable based upon SMSA of residence and found in the main NLSY79 data file named KEYVARS ("Is R's Current Residence in SMSA?"), would give a false impression of mobility in and out of an SMSA for respondents living in the same location for which the SMSA designation was changed between survey years. (See *Appendix 6: Urban-Rural and SMSA-Central City Variables* for further details on the creation of this variable.)

Note that all other SMSA environmental variables for those living in these new areas remain NA, since the *County and City Data Book*, 1972 and 1977 data files did not contain information for these SMSAs. There are 171 NLSY79 respondents who lived in those new SMSAs in 1980 and 2 who lived in the disqualified SMSA in 1979.

Rewrite of 1979–82 Geocode Files

In 1989, work was undertaken to reduce the number of variables provided on the 1979–82 NLSY79 geocode files so that the number and type of data included in these files more closely resembled the geographic data available for the 1983 and subsequent survey years. The previous 1979–82 NLSY79 geocode data file contained 2,245 variables. This number was reduced to 545 variables with county-level and SMSA-level data retained. In addition, four new variables were included in the 1979–82 NLSY79 geocode data. These variables provide data on the "Continuous Unemployment Rate for the Labor Market of Current Residence" for each survey year. This reduction in the number of variables made it possible to better document the geocode data files and to produce codebooks like the ones produced for the main NLSY79 data.

1983–86 Geocode Data File Creation Procedures

The following briefly outlines the procedures used to create the 1983–86 NLSY79 geocode files.

1. State, county, and zip codes are reported by each NLSY79 respondent. Missing information was hand-edited whenever possible. The majority of hand-edits involved the derivation and addition of a zip code. A variable was created to indicate the type of hand-editing that was done on each case.

2. The state, county, and zip codes were then matched with the CRF. For those cases where the NLSY79 state, county, and zip codes matched with a state, county, and zip code from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

3. The NLSY79 file, with SMSA/MSA added when there was a match on all three residence variables, was then merged with the *County and City Data Book*, 1983 data file.

4. For cases missing an SMSA/MSA because there was no match with the CRF, a match was made based on the NLSY79 county and state and the CRF county and state so that an SMSA/MSA could be provided. For those cases, an edit code of 7 was assigned. Respondents living in New England were excluded from this step since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file is the New England County Metropolitan Areas (NECMA) code.

1987 Geocode Data File Creation Procedures

The following briefly outlines the procedures used to create the 1987 NLSY79 geocode files.

1. State, county, and zip codes are reported by each NLSY79 respondent. Missing information was hand-edited whenever possible. The majority of hand-edits involved the derivation and addition of a zip code. A variable was created to indicate the type of hand-editing that was done on each case.

2. The state, county, and zip codes were then matched with the CRF. For those cases where the NLSY79 state, county, and zip codes matched with a state, county, and zip code from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

   In the 1987 NLSY79 data file, if there was not an exact match on the state, county, and zip codes, two additional steps were taken. First, a match was then attempted on the state and zip codes. If the state and zip codes matched, then the county and SMSA/MSA codes from the CRF were added to the respondent's record. This type of match is indicated by a value of 10 on the edit variable. Second, if there was no match on state and zip codes, but there was a match on the zip code only, then the state, county, and SMSA/MSA codes from the CRF were added to the respondent's record. This type of match is indicated by a value of 11 on the edit variable.

   The changes in the matching strategy were made because the zip code was more accurate than the county and state geocodes that were available for the 1987 NLSY79 data. Some mismatching, however, did occur because the zip code was in error rather than the county or state code, but this error rate was smaller than another matching algorithm not requiring case by case hand edits. We are confident that matching by zip code improved the quality of the match.

3. The NLSY79 file, with SMSA/MSA added when there was a match on all three residence variables, was then merged with the *County and City Data Book*, 1983 data file.

4. For cases missing an SMSA/MSA because there was no match with the CRF, a match was made based on the NLSY79 county and state and the CRF county and state so that an SMSA/MSA could be provided for those cases. For those cases, an edit code of 7 was assigned. Respondents living in New England were excluded from this step since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file is the New England County Metropolitan Areas (NECMA) code.

One final caution is that matching by state and zip code or by zip code only may result in a higher moving rate between 1987 and the previous interview year than might actually have occurred. We suspect that some NLSY79 county and state geocodes were not updated if the respondent reported an address change prior to the 1987 interview or the previous interview. If the NLSY79 geocodes were not updated in a previous interview, then there would have been an under-reporting of moving to a new county and/or state in that interview year that would now show up with the 1987 NLSY79 data because of the improved matching algorithm. Due to time and personnel constraints, it was not possible to examine every case that did not initially match on the state, county, and zip codes.

1988 Geocode Data File Creation Procedures

The following briefly outlines the procedures used to create the 1988 geocode files.

1. State, county, and zip codes are reported by each NLSY79 respondent. Missing information was hand-edited wherever possible. More than 1,000 hand-edits were performed. Approximately 56.6% of these involved the derivation and addition of a zip code, while approximately 48.4% involved correction of the state of residence. A variable was created to indicate the type of hand-editing that was done on each case.

2. The state, county, and zip codes were then matched with the CRF. For those cases where the NLSY79 state, county, and zip codes matched with a state, county, and zip code from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

   Beginning in 1987 and again in 1988, if there was not an exact match on the state, county, and zip code, an additional step was taken. A match was attempted on the state and zip codes. If the state and zip code matched, then the county and SMSA/MSA codes from the CRF were added to the respondent's record. This type of match is indicated by a value of 10 on the edit variables. This action involved an additional 1,058 cases.

   This strategy was employed because a match of both zip code and state was more accurate than either variable taken individually. It is probable that some mismatching did occur because the county itself was in error. We are confident however, that by requiring a match on both zip code and state to determine the county, we have improved the quality of the match significantly. In support of this assumption, the cases that were actually hand-edited, produced only approximately 6% with an invalid county. The possibility of zip codes continuing across adjacent counties suggests that this may even be an overestimate of the actual error occurring.

   One additional word of caution applies here. In the 1987 NLSY79 geocode data, if the zip code and state did not match, but the zip code alone matched, the state and county were added to the record. There was a possibility of additional mismatching in cases for which the zip code was incorrect. Because the 1988 procedure required both the zip code and state to match, some cases in which the zip code alone matched, and which were possibly in error in 1987, may have been hand-edited in 1988. This may affect mobility rates between 1987 and 1988 to the extent that those inaccurate zip codes in 1987 have been corrected in the 1988 NLSY79 data file. See the NLSY79 1987 geocode procedures documentation for further discussion of differences in procedures between pre-1987 and post-1987 NLSY79 geocode data.

3. The NLSY79 file, with SMSA/MSA added when there was a match on zip code, county, and state of residence, was then merged with the *County and City Data Book*, 1983 and 1988 data files.

4. For cases missing an SMSA/MSA because there was no match with the CRF, a match was made based on the NLSY79 county and state and the CRF county and state so that an SMSA/MSA could be provided for those cases. For those cases, an edit code of 7 was assigned. Respondents living in New England were excluded from this step, when merging with the 1983 *County and City Data Book* data file, since the SMSA/MSA variable on the 1983 data file for those cases is the New England County Metropolitan Areas (NECMA) code. NECMA residents were <u>not</u> excluded when merging with the 1988 *County and City Data Book* data file. In the 1988 *County and City Data Book* data

file, the MSA/NECMA and the CMSA variables found in the 1983 *County and City Data Book* data file were combined into one 4-digit variable. The addition of a "Record Type" variable in the 1988 *County and City Data Book* (see discussion below regarding variable selection) makes it possible for the user to isolate those living in a NECMA and exclude them from the analysis.

1989 Geocode Data File Creation Procedure

A new procedure was implemented in 1989 as an initial step in verifying the county and state of residence by using address information from the "switch" associated with each area code and exchange. In the hand-editing process for the 1988 Geocode file, reported telephone information was found to be very accurate, even in cases for which some or all of the address information was in error. Thus the telephone information presented itself as a reliable, independent source of verification for the address information. The state and county generated from the phone number are compared to the state and county in the NORC address file for each respondent. All cases in which the telephone information would indicate a different state and/or county from that in the address file are identified through this process. This procedure helps identify respondents with incorrect or inconsistent records. Cases that produced such a non-match were checked for accuracy and hand-edited if necessary.

The following briefly outlines the procedures used to create the 1989 Geocode files.

1. An initial data set was constructed containing state, county, and zip code information for the "switch" which regulates each area code and exchange (the "PHONE" data set).

2. A second data set was constructed containing the state, county, zip code, and telephone information reported by the respondent ("ADDRESS" data) and the state and county information from the CRF for the respondent's reported zip code ("ZIP CODE" data).

3. The state variables from each of these sources were then compared and a "quality of match" variable was computed based upon the extent to which the "PHONE" state, the "ZIP CODE" state and the "ADDRESS" state match. The highest quality match exists if the "PHONE" state, the "ZIP CODE" state and the "ADDRESS" state all match. If a non-match occurred between these state variables, then the geocode information was represented by data matching the "PHONE" information.

4. The state and county established through this matching and verification procedure were then compared to the state and county reported by NORC for each respondent. Cases for which a non-match occurred between states and/or counties were examined individually. These cases were hand-edited if possible.

   From this point, the procedures closely follow those applied in constructing the 1988 Geocode data files, with minor modifications. The CRF matching was based upon state and county only for the purposes of the final matching of information from the *County and City Data Book* data files. As metropolitan statistical area information is based upon county delineations (except in New England), matching on cleaned state and county data should not affect the assignment of respondent MSAs.

5. The state and county were then matched with the CRF. For those cases where the NLSY79 state and county matched with a state and county from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

A word of caution applies. In the 1987 NLSY79 geocode data, if the zip code and state did not match, but the zip code alone matched, the state and county were added to the record. There was a possibility of additional mismatching in cases for which the zip code was incorrect. Because the 1988 procedure required both the zip code and state to match, some cases in which the zip code alone matched, and which were possibly in error in 1987, may have been hand-edited in 1988. This may affect mobility rates between 1987 and 1988 to the extent that those inaccurate zip codes in 1987 have been corrected in the 1988 NLSY79 data file. Additionally, the effect of the 1989 phone verification procedures on the ability to detect errors in the NORC geocode data may also affect mobility rates between 1988 and 1989. See the NLSY79 1987 geocode procedures documentation for further discussion of differences in procedures between pre-1987 and post-1987 NLSY79 geocode data.

6. The NLSY79 file, with SMSA/MSA added when there was a match on county and state of residence, was then merged with the *County and City Data Book*, 1983 and 1988 data files.

7. Respondents living in New England were excluded from this step, when merging with the 1983 *County and City Data Book* data file, since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file for those cases is the New England County Metropolitan Areas (NECMA) code. NECMA residents were <u>not</u> excluded when merging with the 1988 *County and City Data Book* data file. In the 1988 data file, the MSA/NECMA and the CMSA variables found in the 1983 data file were combined into one 4-digit variable. The addition of a "Record Type" variable in the 1988 *County and City Data Book* (see discussion below regarding variable selection) makes it possible for the user to isolate and exclude from analysis those living in a NECMA.

1990 Geocode Data File Creation Procedure

The procedures for the creation of the 1990 Geocode data file are very similar to those applied in creating the 1989 Geocode data file with some small modifications noted below.

1. An initial data set was constructed containing state, county, and zip code information for the "switch" which regulates each area code and exchange (the "PHONE" data set).

2. A second data set was constructed containing the state, county, zip code, and telephone information reported by the respondent ("ADDRESS" data) and the state and county information from the CRF for the respondent's reported zip code ("ZIP CODE" data).

3. The state variables from each of these sources were then compared and a "quality of match" variable was computed based upon the extent to which the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state matched. The highest quality match exists if the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state all match. If a non-match occurred between these state variables then the geocode information was represented by data matching the "PHONE" information.

4. The state and county established through this matching and verification procedure were then compared to 1989 CHRR-edited versions of the state and county of residence reported by each respondent (see R30769. and R30770.). Cases for which a non-match

occurred between states and/or counties were examined individually. These cases were hand-edited whenever possible.

In the 1989 procedure the geocodes established by the phone number were compared to the geocodes received directly from NORC. By using the 1989 CHRR-edited versions of the geocodes for comparison, updates and corrections that were made to the geocodes during the 1989 hand-editing processes were incorporated. This reduced the number of mismatches between the geocode information based upon the current phone number and the respondent-reported geocode information and increased the amount of consistency observed between survey years. The number of cases requiring individual examination was thereby reduced.

From this point, the procedures closely follow those applied in constructing the 1988–89 Geocode data file, with minor modifications. For 1989, CRF matching was based upon state and county only for the purposes of the final matching of information from the *County and City Data Book* data files (see discussion of 1989 Geocode data file creation procedures). A match on state, county, and zip is also required to construct a variable reflecting a respondent's SMSA/non-SMSA residence status for inclusion in the NLSY79 main data file. This match, which was included in the geocode procedures prior to 1989, was done separately for the 1989 release when the new set of initial procedures was instituted. To streamline programming tasks, however, the zip information was reinserted in the CRF matching program for 1990. Therefore, the CRF matching for the 1990 geocode data file was again based upon state, county, and zip code, as it had been prior to 1989.

5. The state, county, and zip code were then matched with the CRF. For those cases where the NLSY79 state, county, and zip matched with a state, county, and zip from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

Researchers should exercise caution when using the geocode data files because several modifications were made after 1987 in the programming procedures that create the files. These include:

a) In the 1987 NLSY79 geocode data, if the zip code and state did not match but the zip code alone matched, the state and county were added to the record. There was a possibility of additional mismatching in cases for which the zip code was incorrect. Because the 1988 procedure required both the zip code and state to match, some cases in which the zip code alone matched, and which were possibly in error in 1987, may have been hand-edited in 1988. This may affect mobility rates between 1987 and 1988 to the extent that those inaccurate zip codes in 1987 have been corrected in the 1988 NLSY79 data file. Additionally, the effect of the 1989 phone verification procedures on the ability to detect errors in the NORC geocode data may also affect mobility rates between 1988 and 1989. See the NLSY79 1987 geocode procedures documentation for further discussion of differences in procedures between pre-1987 and post-1987 NLSY79 geocode data.

b) Residence information is usually collected by NORC interviewers only when there has been a change in that information from the previous interview. In 1990, however, an effort was made to get current information for all respondents. Many of the cases in this current update information also included counties that have been inconclusive (even in case-by-case hand-editing) in previous years. These are

generally cases in which a zip code spans more than one county, and for which valid county data is missing from the respondent's reported residence information. For such cases, the possibility existed in the 1989 (and prior) data that counties assigned based upon such multiple-county zip codes might be in error in a small number of cases. (This would result in the assignment of a county adjacent to the county in which the respondent actually lived.) To the extent that current update information for the county of residence in 1990 shows the assigned county in 1989 to be in error, mobility determinations may be affected. In contrast, using the 1989 CHRR-edited versions of the geocodes for comparison with the current geocode information should improve the accuracy of mobility ratings. This is a more dependable confirmation of past geocode information, eliminating the need to make individual determinations in many cases with multiple-county zip codes as discussed above.

6. The NLSY79 file, with SMSA/MSA added when there was a match on county and state of residence, was then merged with the *County and City Data Book*, 1983 and 1988 data files.

7. Respondents living in New England were excluded from this step, when merging with the 1983 *County and City Data Book* data file, since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file for those cases is the New England County metropolitan Areas (NECMA) code. NECMA residents were <u>not</u> excluded when merging with the 1988 *County and City Data Book* data file. In the 1988 data file, the MSA/NECMA and the CMSA variables found in the 1983 data file were combined into one 4-digit variable. The addition of a "Record Type" variable in the 1988 *County and City Data Book* (see discussion below regarding variable selection) makes it possible for the user to isolate those living in a NECMA and exclude them from the analysis.

1991 Geocode Data File Creation Procedure

The procedures for the creation of the 1991 Geocode data file are very similar to those applied in creating the 1990 Geocode file. The following briefly outlines the procedures used to create the 1991 file.

1. An initial data set was constructed containing state, county, and zip code information for the "switch" which regulates each area code and exchange (the "PHONE" data set).

2. A second data set was constructed containing the state, county, zip code, and telephone information reported by the respondent ("ADDRESS" data) and the state and county information from the CRF for the respondent's reported zip code ("ZIP CODE" data).

3. The state variables from each of these sources were then compared and a "quality of match" variable was computed based upon the extent to which the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state match. The highest quality match exists if the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state all match. If a non-match occurred between these state variables, then the geocode information was represented by the data matching the "PHONE" information.

4. The state and county established through this matching and verification procedure were then compared to 1990 CHRR-edited versions of the state and county of residence reported by each respondent (see R34114. and R34113.). Cases for which a non-match

occurred between states and/or counties were examined individually. These cases were hand-edited whenever possible.

5. The state, county, and zip code were then matched with the CRF. For those cases where the NLSY79 state, county, and zip matched with a state, county, and zip from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

   Because several modifications have been made since 1987 in the programming procedures that create the geocode data files, researchers should exercise caution while doing their analyses. In particular, users should be aware of the following:

   a) In the 1987 NLSY79 geocode data, if the zip code and state did not match but the zip code alone matched, the state and county were added to the record. This resulted in the possibility of additional mismatching in cases for which the zip code was incorrect. Because the 1988 procedure required both the zip code and state to match, some cases in which the zip code alone matched, and which were possibly in error in 1987, may have been hand-edited in 1988. This may affect mobility rates between 1987 and 1988 to the extent that those inaccurate zip codes in 1987 have been corrected in the 1988 NLSY79 data file. Additionally, the effect of the 1989 phone verification procedures on the ability to detect errors in the NORC geocode data may also affect mobility rates between 1988 and 1989. See the NLSY79 1987 geocode procedures documentation for further discussion of differences in procedures between pre-1987 and post-1987 NLSY79 geocode data.

   b) Residence information is usually collected by NORC interviewers only when there has been a change in that information from the previous interview. In 1990, however, an effort was made to get current information for all respondents. Many of the cases with this current update information also included counties that had been inconclusive (even in case-by-case hand-editing) in previous years. These are generally cases in which a zip code spans more than one county and for which valid county data is missing from the respondent's reported residence information. For such cases, the possibility existed in the 1989 (and prior) data that counties assigned based upon such multiple-county zip codes might be in error in a small number of cases. (This would result in the assignment of a county adjacent to the county in which the respondent actually lived.) To the extent that current update information for the county of residence in 1990 shows the assigned county in 1989 to be in error, mobility determinations may be affected. In contrast, using the 1989 CHRR-edited versions of the geocodes for comparison with the current geocode information should improve the accuracy of mobility ratings. This is a more dependable confirmation of past geocode information, eliminating the need to make individual determinations in many cases with multiple-county zip codes as discussed above.

6. The NLSY79 file, with SMSA/MSA added when there was a match on county and state of residence, was then merged with the *County and City Data Book*, 1983 and 1988 data files.

7. Respondents living in New England were excluded from this step, when merging with the 1983 *County and City Data Book* data file, since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file for those cases is the New England County

Metropolitan Areas (NECMA) code. NECMA residents were not excluded when merging with the 1988 *County and City Data Book* data file. In the 1988 data file, the MSA/NECMA and the CMSA variables found in the 1983 data file were combined into one 4-digit variable. The addition of a "Record Type" variable in the 1988 *County and City Data Book* (see discussion below regarding variable selection) makes it possible for the user to isolate those living in a NECMA and exclude them from the analysis.

1992–94 Geocode Data File Creation Procedure

The procedures for the creation of the 1992–94 Geocode data files have been streamlined from those in previous years, particularly in terms of the hand-editing required on individual cases. The following briefly outlines the procedures used to create the 1992–94 Geocode files.

1. A new locator database was created containing the most recent information on each respondent's residence at the time of the survey. Wherever possible, a code was then assigned for each state, county, and country (if applicable). The information in the new locator database was then compared to locator information from the previous interview year. This included an electronic comparison of the character strings entered for street addresses of respondents. If the state, county, and zipcode information matched that from the previous year, and the address strings matched in whole or in significant part (indicating probable typos or keypunch errors), the same state and county geocodes were assigned to a case as were assigned in the previous year. Cases for which a partial or full mismatch occurred, or for which information was missing from any field, were identified during this process and were hand-edited wherever necessary. These cases were then assigned a state and county code based upon the hand-edited data. The procedure of electronic matching of address strings has considerably reduced the number of cases requiring individual hand-editing.

2. Data from each respondent's locator record was then matched by zip code to the state and county from the CRF data file (the "ZIP CODE" data) and by area code and extension to the state and zip from the Local Exchange Routing Guide (LERG) data file (the "PHONE" data). (See discussion of 1991 Geocode data file).

3. The state variables from each of these sources were then compared and a "quality of match" variable was computed based upon the extent to which the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state match. The highest quality match exists if the "PHONE" state, the "ZIP CODE" state, and the "ADDRESS" state all match. Cases in which erroneous or missing zipcode and/or phone information could not be assigned and, in turn, which prevented assignment of state and county geocodes from either the "ZIP CODE" or the "PHONE" data files, were hand-edited as necessary. The matching procedure was then repeated.

4. The state, county, and zip code were then matched with the CRF. For those cases where the NLSY79 state, county, and zip matched with a state, county, and zip from the CRF, the SMSA/MSA from the CRF was added to each respondent's record.

   Researchers are encouraged to use caution during analyses because several modifications were made since 1987 in the programming procedures that create the geocode data files. Specific modifications of note include the following:

a) In the 1987 NLSY79 geocode data, if the zip code and state did not match, but the zip code alone matched, the state and county were added to the record. This created a possibility of additional mismatching in cases for which the zip code was incorrect. Because the 1988 procedure required both the zip code and state to match, some cases in which the zip code alone matched, and which were possibly in error in 1987, may have been hand-edited in 1988. This may affect mobility rates between 1987 and 1988 to the extent that those inaccurate zip codes in 1987 have been corrected in the 1988 NLSY79 data file. Additionally, the effect of the 1989 phone verification procedures on the ability to detect errors in the NORC geocode data may also affect mobility rates between 1988 and 1989. See the NLSY79 1987 geocode procedures documentation for further discussion of differences in procedures between pre-1987 and post-1987 NLSY79 geocode data.

b) Residence information is usually collected by NORC interviewers only when there has been a change in that information from the previous interview. In 1990, however, an effort was made to get current information for all respondents. Many of the cases with this current update information also included counties that have been inconclusive (even in case-by-case hand-editing) in previous years. These are generally cases in which a zip code spans more than one county and for which valid county data is missing from the respondent's reported residence information. For such cases, the possibility existed in the 1989 (and prior) data that counties assigned based upon such multiple-county zip codes might be in error in a small number of cases. (This would result in the assignment of a county adjacent to the county in which the respondent actually lived.) To the extent that current update information for the county of residence in 1990 shows the assigned county in 1989 to be in error, mobility determinations may be affected. In contrast, using the 1989 CHRR-edited versions of the geocodes for comparison with the current geocode information should improve the accuracy of mobility ratings. This is a more dependable confirmation of past geocode information, eliminating the need to make individual determinations in many cases with multiple-county zip codes as discussed above.

c) In creating the 1979–94 geocode data file, the same logical procedures were applied in identifying cases requiring individual examination. However, the automation of the decision rules and procedures to check for and identify such cases resulted in a substantial reduction in the number of hand-edited cases.

5. The NLSY79 file, with SMSA/MSA added when there was a match on county and state of residence, was then merged with the *County and City Data Book*, 1983 and 1988 data files.

6. Respondents living in New England were excluded from this step, when merging with the 1983 *County and City Data Book* data file, since the SMSA/MSA variable on the *County and City Data Book*, 1983 data file for those cases is the New England County Metropolitan Areas (NECMA) code. NECMA residents were not excluded when merging with the 1988 *County and City Data Book* data file. In the 1988 data file, the MSA/NECMA and the CMSA variables found in the 1983 data file were combined into one 4-digit variable. The addition of a "Record Type" variable in the 1988 *County and City Data Book*, (see discussion below regarding variable selection) makes it possible for the user to isolate and exclude from analysis those living in a NECMA.

1996 Geocode Data File Creation Procedure

The procedures for the creation of the 1996 Geocode data file have changed from those used in previous years. The following briefly outlines how the 1996 Geocode file was created.

A new software package, called "Matchmaker for Windows, V2.5" (Matchmaker), was used. Basic geographic information such as latitude and longitude was linked to each respondent's address. This was accomplished by matching address data to information in the Matchmaker database. Matching records were appended with the matching address, coordinates, Census information, and FIPS (Federal Information Processing Standards) codes for state, county, MCD (Minor Civil Division), and MSA (Metropolitan Statistical Area).

Three graduated matching methods were applied, depending on the quality of the address data available.

1. An automated comparison was done between the respondent's address data and the Matchmaker database. Address records with matching street segments were appended with the matching address, coordinates (latitude and longitude values for a specific location), Census information (County, Tract, and Block Group codes for an address), FIPS codes for state, county, MCD, and MSA.

2. For some addresses the procedure outlined in Step #1 failed to produce a match between the respondent's address data and the Matchmaker database. In these cases, individual respondent addresses were temporarily corrected in order to match them to the Matchmaker database. By correcting obvious errors and referring to lists of valid address components, a map display, and commercial maps, a temporary working address was constructed; this was used to assign geocodes to these cases.

   These temporary address corrections were made in a working file to test the improvements in matching to the Matchmaker database. The original address data remained unchanged. Successful address corrections were matched by this method and geocodes assigned accordingly.

3. Addresses unmatched by either of the first two procedures were assigned coordinates and related Census data according to a 5-digit ZIP centroid. A centroid is essentially the mid-point of a ZIP Code area. Centroid matches were made only for addresses that could not been matched by any other means. Addresses with ZIP codes that were no longer current were appended with latitude and longitude coordinates only.

The procedures outlined in steps #2 and #3 approximate the hand-editing process described in previous survey years for records with different degrees of matched address data.

Variable Selection

Variables from the 1988 *County and City Data Book* data file were selected with an eye toward comparability with the 1983 *County and City Data Book* variables. However, some differences do exist between similar variables selected from the 1983 and 1988 *County and City Data Book* data files.

1. The 1983 *County and City Data Book* data file variables for MSA/NECMA and CMSA have been combined into one 4-digit variable in the 1988 *County and City Data Book* data file. Therefore, the 1988 *County and City Data Book* geographic variables

correspond to the 1983 *County and City Data Book* geographic variables in the following manner:

a) The MSA/NECMA codes that existed in the 1983 data file are identical in the 1988 data file.

b) Six MSAs have been added, one MSA has been expanded, and one CMSA has been expanded in the 1988 data file. The MSAs that have been added have their own unique 4-digit code.

c) The 1983 CMSA variable has been recoded with a new unique 4-digit code for each CMSA in the 1988 combined variable. The 1983 PMSA variable has been retained and is identical to the 1988 PMSA variable. Therefore, each CMSA and PMSA is still identifiable in the same manner they were with the separate 1983 CMSA variable.

d) Two 1983 CMSAs were redefined as MSAs in the 1988 data file. These are Kansas City and St. Louis. They have been recoded with their own unique 4-digit code.

e) The addition of a "Record Type" variable in the 1988 data file makes it possible to distinguish separately between MSAs, NECMAs, and CMSAs. This "Record Type" variable classifies cases in the 1988 data file combined MSA/NECMA/CMSA variable according to whether they provide information for the U.S., States, MSAs, NECMAs, CMSAs, or a Nonmetropolitan County. The use of this variable allows the user to exclude any of these groups from the analysis without having to conduct a county-by-county or state-by-state determination of NECMA/non-NECMA status.

2. The population by age variables from the 1988 *County and City Data Book* data file are estimates made for the National Cancer Institute by the Census Bureau. These figures suppress data for counties in which the population is under 20,000. Users should keep this in mind during analysis.

In the absence of updated information from the 1988 *County and City Data Book* data file, the 1983 *County and City Data Book* variables were retained.

1998 Geocode Data File Creation Procedure

A software package called "Matchmaker for Windows, V2.5," used in the creation of the NLSY79 geocode data file, linked address data to information in the Matchmaker database. Matching records were appended with the matching address and FIPS (Federal Information Processing Standards) codes for state and county. Three graduated matching methods were applied depending on the quality of the address data available.

1. An automated match was done between the respondent's locating address data and the Matchmaker database. Address records with matching street segments were appended with the matching state and county FIPS codes.

2. For some addresses the procedure outlined in Step #1 failed to produce a match between the respondent's address data and the Matchmaker database. In these cases individual respondent addresses were temporarily corrected in order to match them to the Matchmaker database. By correcting obvious errors and referring to lists of valid

address components, a map display, and commercial maps, a temporary working address was constructed; this was used to assign geocodes for these cases.

These temporary address corrections were made in a working file to test the improvements in matching to the Matchmaker database. The original address data remained unchanged. Addresses successfully corrected were matched by this method and state and county FIPS codes assigned accordingly.

3. Addresses unmatched by either of the first two procedures were assigned coordinates and related Census data according to a 5-digit ZIP centroid. A centroid is essentially the mid-point of a ZIP code area. Centroid matches were made only for addresses that could not been matched by any other means.

   Users should note that there is some small possibility that a respondent's county may be misassigned using the centroid method in cases where more than one county is represented in a given ZIP code. In these cases, it is possible that a respondent might live in one county but that the center of the ZIP code area is in another county. However, since ZIP codes infrequently cross county lines and less than a quarter of respondents' counties were assigned using the ZIP centroid method, the number of counties incorrectly assigned should be quite small.

Edited variables describing the location of each respondent's residence are created as a result of this matching process. The first two variables, question names "GEO1" and "GEO2", provide the FIPS code for the respondent's county and state of residence. Additionally, "GEO10" provides information about the quality of the respondent's address match—that is, whether the county was assigned based on the respondent-provided address (methods 1 and 2 above) or the ZIP centroid method (method 3 above).

2000 Geocode Data File Creation Procedure

A different software package, Maptitude (V4.2), was used in the creation of the NLSY79 2000 geocode data file. This program links respondent address data to standard geographic information such as the FIPS (Federal Information Processing Standards) codes for state and county. Three graduated matching methods were applied, depending on the quality of the address data available.

1. An automated match was done between the respondent's locating address data and the Maptitude database. Address records with matching street segments were assigned the latitude and longitude of the location. In some cases, addresses had to be cleaned before they could be matched by the Maptitude program. Cleaning involves steps such as standardizing the address format, correcting obvious misspellings, identifying apartment numbers and locating them in the correct field, etc. It does not include any changes that might result in a change in the actual address location.

2. For some addresses, the procedure outlined in Step #1 failed to produce a match between the respondent's address data and the Maptitude database. In these cases, geocode staff used the Maptitude program to locate the correct street. If the street number could be located along this street, the latitude and longitude were assigned. However, some streets in the Maptitude database do not include information about street numbers. If this is the case, the address is manually located in the center of the street. The street is then classified as either a short

street or a long street. Long streets cross Census tract or block group boundaries while short streets do not. As a result, the level of certainty about geographical information is much higher for short streets than for long streets.

3. Addresses unmatched by either of the first two procedures were assigned latitude and longitude coordinates according to a 5-digit zip centroid. A centroid is essentially the midpoint of a zip code area. The geographic information is less certain for respondents located using the zip centroid method.

Researchers can identify the method used to locate the respondent's address by using the variable "GEO10" which provides information about the quality of the geographic match. This variable differentiates between addresses located based on the actual address, in the center of a short street, in the center of a long street, or using the zip centroid method. This variable can be used to determine the level of certainty for the respondent's geographic data.

Supplementary Created Variables

Urban-Rural and SMSA-Central City residence variables

The procedures for creating the Urban-Rural and SMSA residence variables (released in the KEYVARS area of interest) have been modified with the 2000 release. These variables are now created through the Maptitude software as well. If the respondent's residence was located using a street name match (method 2 above) or a zip centroid match (method 3), the MSA and urban/rural variables are further evaluated. For the MSA variable, if the street or zip code falls completely inside or outside the boundaries of the central city, then the respondent is assigned to the appropriate status. If the street or zip code crosses the boundaries of the central city, then the respondent is coded as living in an MSA, with central city status unknown. Similarly, respondents are only assigned to an urban or rural status if their entire street or zip code lies within an urban or rural area. If the street or zip code crosses an urban/rural boundary, the respondent is assigned to an unknown status. For further discussion of these variables, see *Appendix 6: Urban-Rural and SMSA-Central City Variables* in the *NLSY79 Codebook Supplement: Main File 1979–2000*.

Migration History variables

In R19 of the NLSY79 survey, respondents who had moved to a different county or state since the date of last interview were asked to report each address and the dates of each move. The FIPS code for the state and county of each address are included in the R19 Geocode data file. The items collected in the 2000 questionnaire are found in the GEOCODE 2000 area of interest, with question names beginning with "MIGR_". Similar migration histories were collected in several early survey years of the NLSY79.

Missing Data

The missing data values for all items on the geocode data files are -4 and -5. The -5 values indicate a noninterview for a given year. Respondents who have a -4 value in the data for any variables from the *County and City Data Book* fall into the following categories:

1. Respondents who were in the military or who had an APO address;

2. Respondents who were residing outside of the United States;

3. Respondents whose state or county codes could not be determined.

4. Respondents who reside in a county or SMSA/MSA for which the *County and City Data Book* is missing data for that geographic location for that specific item.

5. Respondents who do not reside in an SMSA for any survey year 1979–82 will be missing SMSA level environmental variables for that year.

6. Respondents whose state, county, and zip codes for any survey year 1979–82 do not lead to an unambiguous SMSA designation. This generally applies only to a small number of respondents living in New England.

7. Respondents residing in the New England states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont who did not match on county, state, and zip code on the 1982 or 1983 CRF are coded -4 on all of the metropolitan statistical area variables with NECMA codes for any survey years 1983–87 that they resided in those areas.

8. In the 1988–2000 NLSY79 geocode data file, respondents residing in the New England states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont who did not match on county, state, and zip code on the CRF are coded -4 on all of the 1983 metropolitan statistical area variables with NECMA codes for the survey year 1988–98.

In the 1988–2000 NLSY79 geocode data file, for 1988 metropolitan statistical area variables with NECMA codes for the survey year 1988, respondents living in the New England states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont were not treated any differently than those residing elsewhere. The addition of the "Record Type" variable in the 1988 *County and City Data Book* data file (see discussion on variable selection above) allows the user to designate these cases as missing, and remove them from the analysis, without having to conduct a county-by-county or state-by-state determination of NECMA/non-NECMA status.

Use of the File

Finally, we have a few suggestions concerning the use of these NLSY79 geographic data files. First, the data file and the accompanying documentation should be used in conjunction with the printed versions of the 1972, 1977, 1983, 1988, and 1994 *County and City Data Books* that correspond to each variable desired in order to have complete information regarding variable descriptions and coding idiosyncrasies. Second, the data should not be used in any fashion that would endanger the confidentiality of any sample member. Only those users who have signed a written licensing agreement consenting to protect respondent confidentiality and to other conditions, who agree not to make, or allow to be made, unauthorized copies of the geocode file, and who also agree to indemnify the Center for Human Resource Research for all claims arising from misuse of the file may use these data..