NORC at the UNIVERSITY OF CHICAGO
Innovative solutions in research and technology

Making CONNECTIONS
A Neighborhood Transformation Family Development Initiative

Data Quality in the Retrospective Reporting of Addresses

Prepared for the 65th Annual AAPOR Conference
Chicago, IL
May 2010

Presented by:
Kate Bachtell
Virginia Tangel
Michael Latterner
Ned English
Catherine Haggerty

National Opinion Research Center (NORC)
at the University of Chicago
55 East Monroe St, Suite 2000
Chicago, IL 60603
312-759-4000

Funded by a grant from
The Annie E. Casey Foundation

May, 2010

ABSTRACT

While tracking the movement of respondents has always been crucial for panel studies, the increasing popularity of geographic analyses has furthered the demand for both accurate and systematic address collection. This paper advances the existing literature on retrospective reporting in surveys by focusing on the collection of respondents' past addresses. It features data from the third wave of Making Connections, a ten-year, neighborhood-based survey funded by the Annie E. Casey Foundation. The wave three questionnaire featured a new series that solicited a detailed history of respondents' movement during the previous three years. Recovering previous addresses presented challenges beyond those typically associated with the systematic recording of physical addresses because recalling information from the past is inevitably more difficult than describing the present (Kennickell and Starr-McCluer 1997).

We designed an experiment to test different methods of maximizing data quality in the retrospective address series collected in 2009 in White Center (Seattle), Washington. Households were randomly assigned to two treatment groups. Addresses collected from the first group underwent administrative data cleaning (using Google Maps, etc.) while those in the second group received intensive follow-up calls by a team of field interviewers and managers. We compare the results of these treatments to the original data collected in White Center and investigate the efficacy of each method for producing addresses that can be successfully translated into geographic coordinates for spatial analyses. We find that the retrieval effort – while more costly to execute – was far more successful in returning 'geocodable' addresses. This supports the argument that successful collection of retrospective addresses depends on an interactive process between the interviewer and respondent involving a variety of probing techniques. Our findings may inform improved methodologies for collecting retrospective data in both panel and cross-sectional surveys.

INTRODUCTION

Data quality in the retrospective reporting of data by respondents has always been a concern in survey research (Jobe, Tourangeau, and Smith 1993). Retrospective reporting places considerable demands on respondents, requiring them to assess whether or not a particular event occurred, recall the timing of the event, and, in some cases, the sequence in which the event occurred. Because of this laborious cognitive process, memory retrieval can be bolstered when substantively contextualized in the larger survey (Wu, Martin, and Long 2001:522, 523). Moreover, previous studies suggest that the accuracy of the self-reports *decreases* as the frequency of similar events increases or as the events become "complex, irregular, or variable in nature" (Wu, Martin, and Long 2001: 523). Respondents are believed to have particular difficulty remembering specific dates of 'unrehearsed' events (as opposed to 'rehearsed' or 'landmark' dates like anniversaries and birthdays) because they must be reconstructed during the interview process using a variety of heuristics. As a result, the quality of 'unrehearsed' data decreases as the duration recall length increases (Wu, Martin, and Long 2001:552).

While the use of administrative records and aided recall procedures have been the two principal methods used in survey research to facilitate the collection of self-reported retrospective data, they have different effects on memory errors. "The use of records generally controls for overstatements due to compression-of-time errors, but has no or only a small effect on errors of omission. Aided recall, on the contrary, reduces the number of events that are omitted, but does not reduce (and may even increase) telescoping effects[1]" (Sudman and Bradburn 1973:805). Thus, errors of omission and errors of overstating events within a particular timeframe are both common problems affecting data quality in the retrospective reporting of events. Addressing the problem of timing recall in the reporting of events, Belli, Shay and Stafford (2001) used an experimental design to find that event history calendars produced higher quality retrospective data on moving than traditional question lists. While many studies have examined the accuracy of data reported by respondents (by means of external validation), only a few have explored the particular challenges associated with the retrospective collection of

[1] Telescoping refers to the tendency for respondents to over-report behaviors or events by including those that occurred outside of the requested timeframe. Telescoping errors typically occur when a respondent overestimates the recency of events in an effort to thoroughly perform the task asked of him/her.

addresses (i.e. Belli, Shay and Stafford 2001) and none to our knowledge have examined the geographic precision of the addresses collected. Accurate address information is critical to the execution of surveys and carries tremendous potential for geographic analyses of behavioral and attitudinal measures.

Our data are derived from the third wave of the *Making Connections* study, sponsored by the Annie E. Casey Foundation. Respondents were asked to list each housing unit in which they resided since November, 2005, as well as the month and year during which they moved there. A simple question list was incorporated in a paper and pencil (PAPI) questionnaire to record this information, preceded by a series of questions about neighborhood conditions. The use of question lists has brought concerns about the validity of retrospective event data (Suchman and Jordan 1990) but have been proven to be effective when strict adherence to standardized language is deemphasized (Fowler 1995).

This paper seeks to assess which techniques are best-suited to elicit and record the highest quality data regarding respondents' previous places of residence. We briefly investigate the role of retrospective address reporting in panel studies like *Making Connections* with an ear to the interests of both data collection agencies and geographic analysts. We weigh the logistical and cost considerations associated with various data cleaning techniques and approach 'quality' as the ability to translate self-reported addresses into precise geographic coordinates.

Survey Background

The *Making Connections* survey is conducted in 10 low-income neighborhoods across the U.S. The neighborhoods are located in the metropolitan areas of Des Moines, IA; Indianapolis, IN; Denver, CO; San Antonio, TX; Seattle, WA; Milwaukee, WI; Oakland, CA; Hartford, CT; Providence, RI; and Louisville, KY. The survey is part of a larger initiative funded by the Annie E. Casey Foundation aimed to assess the needs of families and children and to foster supportive communities that meet those needs. A number of organizations contributed to the design, implementation, and analysis of this research: the Annie E. Casey Foundation, Local Management Entities (LMEs)[2], National Opinion Research Center (NORC), Urban

[2] An LME is defined as "a consortium of people and organizations with data-related expertise or interest in each Making Connections site…" LMEs contribute to the Making Connections initiative in three areas: "development of

Institute (UI), research advisors from Chapin Hall at the University of Chicago and Case Western Reserve University, and representatives from each of the *Making Connections* sites.

Baseline survey data were gathered between 2002 and 2004 in the ten sites listed above and a first follow-up effort ('Wave 2') was completed between 2005 and 2007 in each site. Between 2008 and 2011 NORC will complete a second round of follow-up interviews ('Wave 3') in seven of the ten sites. Interviews for the *Making Connections* neighborhood surveys are executed using a paper and pencil questionnaire that is then keyed into a computer-assisted data entry system (CADE). See Table 1 for a summary of the Baseline (Wave 1), Wave 2, and Wave 3 data collection.

Table 1. Summary of Baseline and Follow-Up Data Collection Efforts

Effort	Data Collection Period	Mode
Wave 1	2002-2003, Louisville in 2004	In person
Wave 2	2005-2006, Louisville in 2007	In person and by telephone
Wave 3	2008-2009, Louisville scheduled for 2011	In person and by telephone

Overview of Making Connections Sample

The Making Connections study design is unique in that it combines both cross-sectional and longitudinal (panel) methodologies. In each wave, NORC employs area probability sampling techniques to select a random set of addresses to represent each target neighborhood. In waves 2 and 3, interviewers re-visited these sampled addresses in person or by telephone with the goal of collecting data with the current occupants. Many times, the occupants have not changed (we call these 'stayer' households). Other times, new people have moved in (we call these 'new households'). We also subsample new addresses at the start of each follow-up effort to include buildings that have been constructed or renovated since the previous wave ('fresh cases'). This methodology yields a cross-sectional snapshot of neighborhood residents at different points in time.

outcomes, measures, and strategies to achieve results, local and cross-site evaluation, and creating a learning community" (www.aecf.org, 2004).

The study is also longitudinal in that 1) we re-interview families that remain at sampled addresses within Making Connections neighborhoods and 2) we track families with children that move to a *new* address, be it inside or outside of the neighborhood. To facilitate these efforts we collect address information for previous wave household members using a number of different methods. Approximately halfway through the three-year interval between waves, we mail postcards to previous participants asking them to notify us if they have changed or expect to change their address or phone number. A few months prior to the start of data collection in each site a team of field locators review this information in combination with the contact information gathered during previous rounds of the survey. They perform individual and batch searches using Accurint (www.accurint.com) and other search engines to gather updated information for all households with children and for a subset of adult-only households. In addition to these efforts aimed at locating *people*, a team of field locators and survey methodologists visit housing units in the neighborhood to validate *addresses*: they identify missed housing units and vacant or demolished buildings. The data from each of these activities is processed to inform the materials supplied to interviewers for the purposes of locating and interviewing sampled households.

Mobility

Movement among *Making Connections* households is frequent, as is commonly observed in populations with high percentages of lower income minority groups and renters. Data from the 2008 Current Population Survey suggest that the percentage of individuals who had moved in the last year was over five times higher among occupants of rental units versus those in owner-occupied units. In administering wave 2 of *Making Connections* NORC found that over half (55%) of all households and 61% of households with children[3] had moved[4] since wave 1 (see Table 2). These moves tended to be short-distance: the median distance for all completed child 'movers' at wave 2 was 2.4 miles (unweighted) and 32.6% remained in the target *Making*

[3] As defined in wave 1.

[4] By 'move' we do not mean that an entire household relocated from their wave 1 housing unit as an entity. Rather, for wave 1 households with children, we define 'move' as the focal child residing in a different housing unit at wave 2. For wave 1 households without children, we base the movement definition on the wave 1 respondent. This nuanced definition of movement is in keeping with *Making Connections'* hybrid cross-sectional and longitudinal design.

Connections neighborhoods. The findings in sites like Louisville reveal a large reliance on public housing among sample members. The demolition of several subsidized developments contributed to a movement rate of 67% among all households and 77% among households with children. Additionally, individual sample members – particularly children – sometimes returned to previous addresses after moving away. These characteristics heighten the need to maintain thorough records of respondents' whereabouts in between data collection waves.

Table 2. *Household Mobility Between Waves 1 and 2*

Site	Wave 1 Total Households	% of Total Households that Moved	Wave 1 Households w/ Children	% of Households w/ Children that Moved
Denver	703	57.2%	316	64.6%
Des Moines	700	50.4%	358	55.0%
Hartford	442	60.9%	307	66.1%
Indianapolis	700	55.6%	349	68.2%
Louisville	704	67.3%	285	77.9%
Milwaukee	697	59.7%	422	66.1%
Oakland	697	58.8%	359	57.1%
Providence	700	54.9%	429	57.1%
San Antonio	700	43.6%	434	53.5%
White Center	700	47.8%	349	53.3%
Total	6,743	55.4%	3,608	61.3%

The Mobility Series

With the wave 2 findings in mind, the *Making Connections* research team decided to include a series of questions in the wave 3 main questionnaire that would capture additional movement occurring in between rounds that may have evaded our other locating efforts. The question design was inspired by items appearing in the Interim Evaluation survey for the Moving to Opportunity for Fair Housing Demonstration Program, a project sponsored by the U.S.

Department of Housing and Urban Development (HUD). The basic goal was to enumerate every address at which the respondent had lived for at least one month since "Thanksgiving, November 2005." We chose November because it fell at roughly the midpoint of the wave 2 data collection calendar and referenced Thanksgiving in order to tap into more autobiographical memories attached to public holidays (Belli et al 2001). To begin the series, field interviewers asked respondents when they had moved to their current address (see Figure 1).

Figure 1: *Beginning of Mobility Section from Wave 3 Making Connections Paper and Pencil Interview (PAPI) Survey*

Next I want to ask you some questions about the different places you have lived since Thanksgiving, November, 2005. This includes any place you lived or stayed for one month or more.

I have recorded that you currently live here at [**1 BEST ADDRESS FROM FACESHEET**].

1.9 In what month and year did you move to this address?

_____ MONTH _____ YEAR

RESPONDENT WAS BORN AND RAISED AT THIS
ADDRESS AND HAS NEVER MOVED..... 1 → **GO TO SEGMENT 2**
DON'T KNOW..... DK
REFUSED..... REF

INTERVIEWER QUESTION: IS THE MONTH AND YEAR AT QUESTION 1.9 EARLIER THAN NOVEMBER 2005 (11/05)?.

YES..... 1 → **GO TO 1.12**
NO..... 2

If the respondent indicated that he or she had moved to their current address *before* November 2005, they were skipped to the next section. If they had moved *during or after* November 2005, the interviewer collected their previous address (see Figure 2).

Figure 2: *Latter Part of Mobility Section from Wave 3 Making Connections Paper and Pencil Interview (PAPI) Survey*

1.10 Where did you live before moving to this address?

ADDRESS: _____ UNIT#: _____

CITY: _____

STATE: _____ ZIP: _____

RESPONDENT WAS HOMELESS 1
DON'T KNOW..... DK
REFUSED..... REF

The interviewer then repeated the 'move date' question, this time asking, "In what month and year did you move to *that* address?" (emphasis added). The interviewer again evaluated this date to determine if it fell during or after November 2005. If so, they continued to repeat the series up to five times, each time asking for the respondent's previous address and move date in reverse chronological order.[5] The resulting data include up to six addresses and seven move dates per household. In this paper we explore the data quality challenges associated with the retrospective reporting of addresses as a first attempt to evaluate the methodological richness of this series.

Training

Because the mobility series was not included in previous waves of the survey and would be somewhat taxing for respondents in terms of both time and cognition, we included it in a segment of the wave 3 interviewer training devoted to difficult sections of the questionnaire. In our Des Moines, Indianapolis, Denver, San Antonio, and White Center (Seattle) trainings, we focused on preparing the interviewers to navigate through the mobility series and follow the skip instructions as intended. We rehearsed administering the questions with respondents who had not moved since Thanksgiving of 2005 as well as with those had moved one or more times. We drew an image of a timeline beginning at 2005 and ending at the current date, and explained that

[5] Copies of the full wave 3 questionnaire are available upon request. Please send inquiries to Kate Bachtell at bachtell-kate@norc.org.

the questions were designed to trace respondents' movement back far enough to predate this three-year window. See Appendix 1 for an excerpt of the relevant training materials. We also provided interviewers with guidelines for the systematic recording of all contact information based on United States Postal Service (USPS) standards (see Appendix 2).

Upon reviewing the data collected in the mobility series from the first five sites, we discovered that additional efforts would be necessary in order to prepare the addresses for geographic analysis. We created a new module for the Providence interviewer training that dealt exclusively with the mobility series addresses. We again covered the flow of the questions and goal of capturing movement over the previous three years, but added a discussion of probing techniques inspired by 1) the existing literature on retrospective reporting and 2) best practices identified by our field management staff (included in Appendix 3).

DATA AND MEASURES

Perhaps not surprisingly, only a very small number of respondents (1%) in Des Moines, Indianapolis, Denver, San Antonio, and White Center (Seattle) – referred to as the 'five sites' hereafter – indicated that they had been born and raised at their current address. In other words, they could not report a date on which they moved to their current residence. This group contributed to a majority of respondents (57%) who reported no previous addresses in the three-year window of interest. Roughly a quarter of respondents (26%) provided one address prior to their current addresses, 13% provided two previous addresses, and a final 5% provided three or more addresses. As one would expect, the distribution narrows as the frequency of movement increases. A total of 2,782 addresses were collected from the 1,771 respondents who had moved at least once since November of 2005 across the five sites.

While it is possible to geocode addresses using only a city name or intersection, this incomplete information presents a problem when attempting to allocate the resulting coordinates to Census geographies. The boundaries of Census block groups and other fine-level geographies are typically defined by streets, making it difficult to locate those intersections that fall along a boundary within a particular Census area. For the purposes of tracking mobility with a high

level of precision, it was necessary to geocode the self-reported addresses to the street address position or ZIP+4 centroid. Thus an address' ability or failure to be geocoded to the street level is the primary (dichotomous) dependent variable used in our analysis.

FINDINGS

After examining the quantity of movement reported by respondents we set out to investigate the *quality* of the addresses recorded. Figure 3 outlines the process through which we prepared for and executed an experiment to evaluate different post-interview interventions intended to increase the percentage of addresses that could be translated into precise geographic coordinates. We employed this arguably strict criterion to address the analytic goal that had inspired the adaptation of the mobility series for wave 3: to fill in the gaps between data collection waves and trace detailed patterns of household mobility.

Figure 3. *Intervention Plan for the Retrospective Addresses*

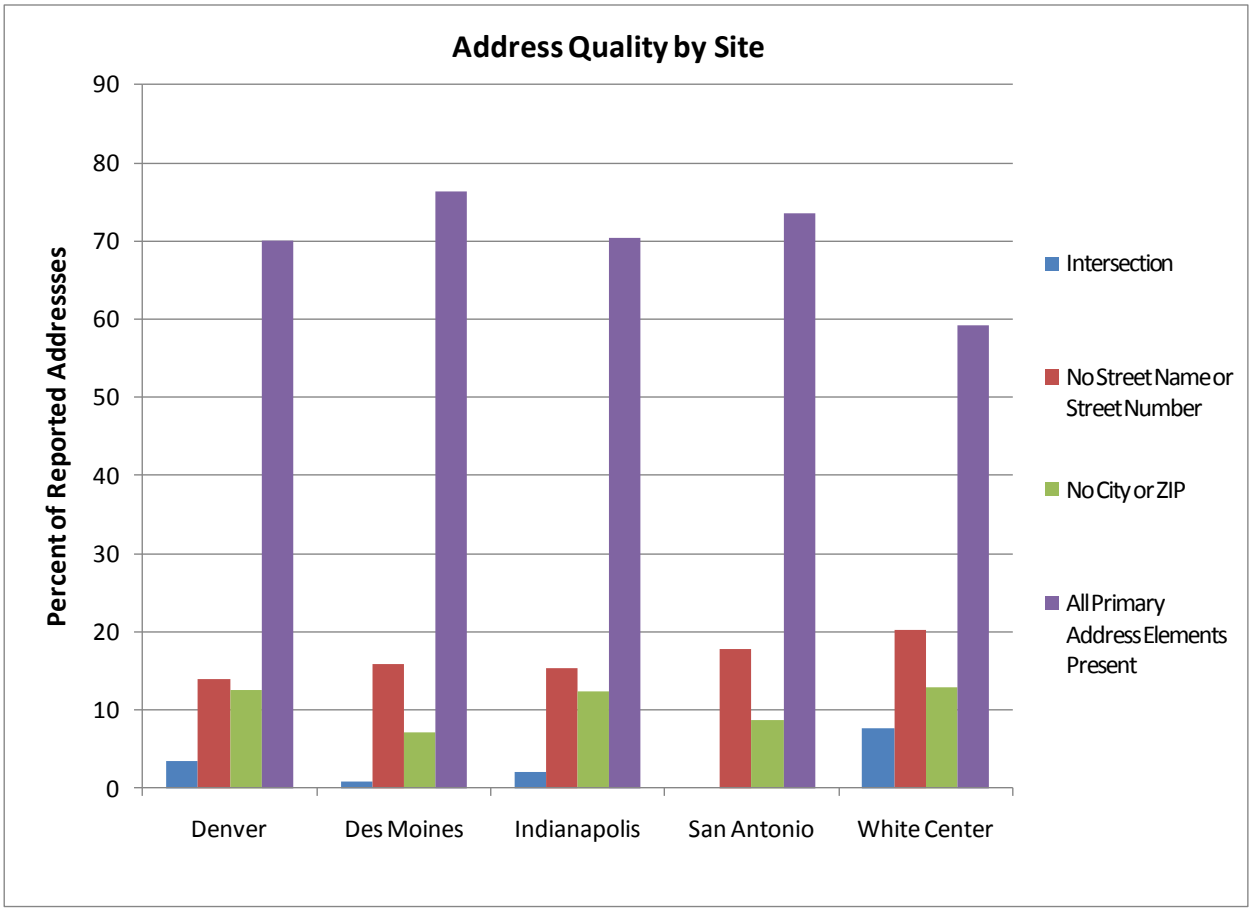
Step 1: Interviewer asks respondent for up to 6 previous addresses (winter – spring 2009)
Step 2: Use *MapMarker Plus* to geocode data and identify ungeocodable addresses (spring 2010)
Step 3: Assign cases to 2 groups (administrative cleaning and retrieval)
Step 4: Execute cleaning treatments
Step 5: Use *MapMarker Plus* to geocode cleaned addresses

Initial Assessment of Data Quality for Five Sites:

Prior to any data cleaning, we evaluated the addresses collected from the mobility series in the five sites. We created an address-level data set in which each prior address reported by the respondent occupied a row. Because we were interested in mapping intricate movement at the street level (recall the findings stated above indicating that mobility among *Making Connections* households tends to be fairly frequent and of close proximity), we wanted to first examine the

extent to which the addresses provided by respondents contained all elements of a complete address: street name and number, city, state, and zip code. Nearly 70% met this criterion across all five sites. Des Moines had the highest percentage of complete addresses (76 %) while the figure was below 60% in White Center (Seattle). We then looked for patterns contributing to the shortcomings of the incomplete addresses and categorized each record along a scale of data quality. 2.8% of the addresses revealed probing efforts by the field interviewer to recover a 'good' address; they recorded the intersection closest to the respondent's home. Some analysts might be content to extrapolate these data to the centroid of a Census block, but they did not meet our criteria for 'immediate geocodability' and so we allocated them to the poorest quality category. 16.5% of the addresses lacked a street name or street number and about 11% were missing a city or zip code. The differences in address quality by site are displayed in Figure 4.

Figure 4. *Initial Assessment of Address Quality for Five Sites*



Geocoding Results for Five Sites

Next, we attempted to geocode the addresses collected in the five sites using MapInfo's *MapMarker Plus* Geocoding software. We first set the preferences to enable automatic geocoding, an approach that involves no interim data cleaning and yields geographic coordinates only for easily-recognizable addresses. Roughly 58% of the addresses (n=1,622) were successfully geocoded. Due to our strict criteria for street addresses, no coordinates were returned for those addresses that contained an intersection (n=79) or lacked a street name or number altogether (n=458). Figure 5 displays examples of these 'ungeocodable' addresses that have been edited to protect respondents' identities. 52% of the 301 addresses with an unspecified city or zip code successfully geocoded, and the success rate among complete addresses -- those with a street name and number, city, state and zip code (n=1,944) -- was over 75%. The sources of error in the complete group include misspellings, P.O. boxes, missing or faulty directionals (i.e. "NW" instead of "SW"), and other miscellaneous problems.

Figure 5. *Examples of ungeocodable addresses (edited to protect confidentiality)*

Address	City	State	Zip
51st & Harlem Ave	Seattle		
Walgreens at 31st & Dearborn Ave	Des Moines	IA	50311
	Indianapolis	IN	46225
PO Box 547	Denver	CO	80206

We then utilized *MapMarker Plus'* interactive geocoding capability. This feature identifies problematic addresses and allows the user to evaluate suggested corrections. For example, where an initial address listed "101 Oak St," *MapMarker* would ask, "Do you mean '101 N Oak St?'" The software also allows the analyst to identify instances in which one or more components of a complete address (as defined for our analysis), such as a street name or zip code, has been updated by the United States Postal Service. We typically employ the interactive mode as a default because it introduces a layer of cleaning that is missed in the automatic mode and produces more mappable results. The tradeoff is an increase in labor for the analyst. In this instance the interactive mode yielded a 10% increase over the automatic mode return of

geocoded addresses for roughly four hours of review time by a survey methodologist. The main area of improvement resided in the complete address group: nearly 88% were successfully geocoded versus 75% after the automatic mode run. There was also a 9% increase in the success rate for those addresses missing a city or zip code (from 52% to just under 62% in the automatic and interactive modes, respectively).

Table 3. *Interactive Geocoding Results for Five Sites*

Address Quality	Total Addresses	Successfully Geocoded
Intersection	79	0.0%
No Street Name or Street Number	458	0.0%
No City or ZIP	301	61.8%
All Primary Address Elements Present	1,944	87.9%
Total	2,782	68.1%

Note: Percentages and cell counts in this and the proceeding tables are unweighted.

Hierarchical Logistic Regression (HLM) Results for Five Sites

Using the dichotomous variable measuring whether or not an address was interactively geocodable, we modeled a hierarchical logistic regression model to predict—at a baseline—what characteristics of respondents and the address date could accurately predict whether or not an individual address would be geocodable.

Independent variables:

Because we were working with an address-level dataset, we built a hierarchical model to nest addresses in individual respondents, thereby controlling for variation within respondents. In addition to this structural rationale, aggregate individual-level variables were assessed in the mixed-effects model on the theoretical grounds that demographic and psychological factors play roles in determining the quality of a respondent's ability to retrospectively report 'good' address(es).

Individual-level measures:

We created a dummy variable indicating whether or not the respondent is a high school graduate to provide a rough measure of socioeconomic status. We controlled for a number of demographic characteristics, including the respondent's sex (where male=1), age, and the native language of the respondent (measured by the language in which their questionnaire was administered). Also included as an independent variable is a variable used to classify each household's mobility between waves of data collection, called MOVEID. Those who stayed at their sampled addresses between waves 2 and 3 ('stayers') served as the reference category. Finally, we included the number of prior addresses given by individual respondents, as well as the position of a particular address within the mobility address series (with later positions indicating older addresses) as predictors of geocodable addresses.

Random effects

Due to the nature of the data set (as an address-level file), the grouping variable was the individual respondent, where each respondent had an individual intercept.

Methodology

Using the 'xtmelogit' routine in STATA 10, a series of mixed-effects logistic regression models were estimated. The initial random intercept model assumes that each respondent occupies a specific regression line, and that their regression lines are parallel (Rabe-Hesketh and Skrondal 2008). Adding a random coefficient of field interviewer (to Model 3) introduces field interviewer-specific slopes to the model, thus relaxing the assumption that respondents' coefficients are parallel (Rabe-Hesketh and Skrondal 2008:155).

Three separate models were estimated. The first shows variation in interactively geocodable addresses within individual respondents. Model 2 introduces individual-level fixed-effects predictors to the model. Model 3 includes the same fixed-effects predictors as Model 2, but also considers a random coefficient for the field interviewer.

Results

A total of 1,644 respondents (who provided 2,586 total addresses) were used for the analysis.[6] Referencing Table 4, Model 2 shows that education (high school graduate versus non-high school graduate) is not a significant predictor of the ability of the address to be geocodable. Demographically, respondents who self-identified as black or African American are 1.4 times more likely than white respondents to report geocodable addresses[7]. Males are .7 times less likely to report automatically geocodable addresses than females. Households with children that were classified as 'movers' (having moved from wave 2 to wave 3) are 1.7 times more likely to report a geocodable address than those respondents who stayed at their sampled address from wave 2 to wave 3. New residents of sampled addresses ('new households') are significantly *less likely* to report automatically geocodable addresses than those wave 2 respondents who remained at their original addresses. For each additional address that a respondent provides, their reported addresses are .8 times less likely to be geocodable ($p<.001$). Interestingly, the sequence of the address in the respondent's memory was not a significant predictor as to whether or not the address itself would be geocodable. Lastly, questionnaires administered in English are 1.8 times more likely to produce geocodable addresses than questionnaires administered in Spanish or Vietnamese.

The introduction of the random coefficient for the field interviewer in Model 3 indicates no further explanatory power offered by that variable (likelihood ratio chi-square=1.73 vs. likelihood ratio chi-square in Model 2, $p<.189$), indicating that the bulk of variation in predicting geocodable addresses lies within individual respondents.

Thus, black respondents, respondents who have moved from wave 2 to wave 3, and respondents who elected to have the questionnaire administered in English were significantly more likely to provide geocodable addresses) than others.[8] With these models of predictors of

[6] Respondents and their associated addresses were dropped from the analysis that had missing values on any of the predictor variables (total n=196 addresses).

[7] Asians, Native Hawaiians/Pacific Islanders, American Indians, and 'Other' races were not significantly more or less likely to report geocodable addresses than whites.

[8] We speculate that interviewers may be slightly better equipped to report higher quality addresses when interviewing 'movers' because the 'face sheet' (page containing case-specific locating information) provided for these cases would display at least one previous address. This merits further research. In addition, it should be noted that the *Making Connections* sample is not nationally representative.

geocodable addresses, we are best able to contextualize an experiment to test different post-interview methods for improving the quality of non-geocodable addresses.

Table 4. *Multi-Level Logit Models Predicting Geocodable Addresses*

Independent Variables	**Model 1**	**Model 2**	**Model 3**
	Odds ratio	***Odds Ratio***	***Odds Ratio***
High school graduate		1.1493	1.1538
		(.13)	(.13)
Race: Black		1.3937	1.3976
		(.20)	(.20)
MOVEID = Mover		1.7286***	1.7455***
		(.27)	(.28)
MOVEID = New W3 Household		.6100***	.6155***
		(.08)	(.08)
MOVEID = Fresh case		.6468	.6530
		(.26)	(.26)
Sex		.7491**	.7452**
		(.10)	(.10)
Age		.9954	.9958
		(.00)	(.00)
English questionnaire		1.8430***	1.8214***
		(.34)	(.34)
Prior address #1		2.3086	2.2721
		(1.78)	(1.75)
Prior address #2		1.3740	1.3498
		(1.05)	(1.03)
Prior address #3		.8290	.8096
		(.64)	(.62)
Prior address #4		1.4893	1.4686
		(1.20)	(1.19)
Prior address #5		1.7583	1.7535
		(1.54)	(1.54)
Number of addresses		.8252***	.8265***
		(.05)	(.05)
Random effects			
Field interviewer			.0057
			(.00)
Within-individual variation	.9631	.9132	.8403
	(.11)	(.12)	(.14)
Goodness of Fit Statistics			
Log likelihood	-1582	-1490	-1489
Wald X2	.	140.90***	138.95***
N	2586	2586	2586

Source: Annie E. Casey Foundation, *Making Connections* Study, Wave 3
Numbers included in parentheses are standard errors
*p<.05 **p<.01 ***p<.001

EXPERIMENT WITH WHITE CENTER DATA

To evaluate our options for improving the quality of addresses collected in the mobility series, we devised an experiment for those addresses that were discovered to be ungeocodable. Because White Center (Seattle) had the highest percentage of ungeocodable addresses (43%) and had completed data collection most recently (spring 2009), we randomly assigned the cases with 'bad' addresses (n=223) into two treatment groups.[9] The first group would undergo administrative cleaning by a team of research assistants who had access to free online resources like Google Maps and MapQuest, as well as a repository of the addresses collected from previous locating attempts for each case (see the overview of the *Making Connections* sample above for an explanation of these efforts). The second group would be subjected to data retrieval by a small team of experienced field interviewers and field managers. The plan was to execute these treatments and compare the cleaned addresses from each group to the pre-treatment findings from the five sites.

Administrative Cleaning

106 of the 'bad' addresses collected in White Center were sent for data cleaning by NORC research assistants. They filled in missing cities, states, and zip codes using free online mapping applications (mainly Google Maps) where possible based on the availability of other information (i.e. street address). Some respondents had reported the name of an establishment like an apartment complex, shelter, or correctional facility without a street address. The team used Google's search capabilities to impute the street name and number for these entries. Other records contained a complete address that was not recognized by *MapMarker Plus*; if the team could not identify and correct a spelling or directional error they inserted a nearby intersection.[10] In other instances, only a street name or city and state had been provided, leaving few options for

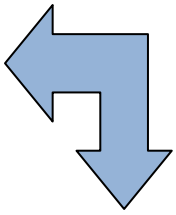
[9] One could argue that it would have been more ideal for the experimental design to randomize selection at the address level rather than the case (respondent) level, but this would present logistical challenges for the phone retrieval group. We did not want to risk losing the opportunity to improve one address with the aid of a respondent while retrieving others because the former had been assigned to the administrative cleaning group. Additionally, the impact of colinearity when analyzing multiple addresses provided by the same respondent should be comparable across both groups.

[10] Because intersections were allocated to the poorest end of our data quality schema, we later applied a condition to preserve the initial data in the event that the cleaned data registered a decrease in quality. This condition is reflected in the analysis presented in later sections.

data cleaning. We retained the original information recorded in these instances. Examples of these scenarios are included in Figure 6. Again, the data have been edited to obscure respondent's true histories.

Figure 6. *Examples of Administratively-Cleaned Addresses (edited to protect confidentiality)*

Original Data			
Address	**City**	**State**	**Zip**
Fox Hollow	White Center	WA	98476
Broadway Shelter on Clark St	Seattle	WA	
24703 101st Pl SE	Seattle	WA	
Near C St close to Parker Ave	Poulsbo	WA	98476



Cleaned Data			
Address	**City**	**State**	**Zip**
Fox Hollow	White Center	WA	98476
501 N Clark St	Seattle	WA	98148
24703 SE 101st St	Seattle	WA	97248
1701 C St SE & 102 1st St SE	Poulsbo	WA	98476

The most pronounced advantage to the administrative cleaning technique is the low cost involved. The cleaning was performed in less than 8 hours and incurred no additional charges from address verification searches.

Phone Retrieval

The other treatment group (n=117) received a personal follow-up (or attempts to follow-up) from a team of experienced field interviewers and field managers. The retrieval team contacted respondents by telephone[11] and attempted to 1) retrieve omitted address elements (i.e. street number) and 2) correct unrecognizable addresses. They recited each 'bad' address to the

[11] While about half of the wave 3 interviews were completed in-person, face-to-face retrieval was not possible due to budget and time constraints.

respondent and explained that we needed help clarifying exactly where he or she lived. A common probing technique was to ask if the respondent could recall the name of their apartment building, which the team then entered into a Google search and recorded as a specific street address. The team also used Accurint – a suite of search tools available to government, research, social service, and revenue agencies – to further evaluate and clean the ungeocodable addresses and to gather further contact information when needed. Accurint returns phone numbers, names, and mailing addresses associated with individuals along with the period of association, making it possible for the retrieval team to discern matches among the list of 'bad' addresses. They used the Accurint addresses to correct spelling and directional errors in the mobility series data and to fill in missing components (i.e. inserting a street name if the street number, city and state matched). This was done cautiously so as to avoid speculation and the risk of over-reporting 'good' addresses which had been collected at a different stage in the mobility series. The drawback associated with using Accurint is that it can be expensive when working with large data sets; the cost for individual searches ranges from about $.35 to $1.80 for returned information. Labor costs for the phone retrieval effort summed to roughly 16 hours.

Treatment Results from White Center

Table 5 reports the categorization of data quality for White Center addresses before and after treatment. The difference in resources available to the two treatment teams is pronounced in the decrease in the number of intersections and missing street names/street numbers. The retrieval team successfully cleaned all but one of the intersections while 17% remained after treatment in the administrative group. Similarly, only five of the missing street names or numbers were filled in by the administrative group while the retrieval team narrowed this group by 35% (from 59 to 18). As a result of these successes the retrieval treatment yielded a much higher return of complete addresses, from a baseline of just 25 to 96 after cleaning (the latter representing 82% of the total addresses assigned to the retrieval group).

Table 5. *Initial Assessment of Data Quality for White Center before and after Treatment*

Address Quality	Administrative		Retrieval	
	Before	After	Before	After
Intersection	18.9%	17.0%	17.1%	0.9%
No Street Name or Street Number	44.3%	39.6%	50.4%	15.4%
No City or ZIP	14.2%	10.4%	11.1%	1.7%
All Primary Address Elements Present	22.6%	33.0%	21.4%	82.1%

Note: The ns for the administrative group and retrieval groups are 106 and 117, respectively, for a total of 223 observations.

Interactive Geocoding Results for White Center addresses

The disproportionate returns from the retrieval effort held up upon re-running the post-treatment data through *MapMarker Plus* in the interactive mode. Nearly 71% of the cleaned addresses in this group were successfully geocoded, versus only 16% in the administrative group. Considering that *none* of the addresses assigned to either group were geocodable prior to treatment, both methods predictably yielded an improvement in data quality. Yet the 55% difference in magnitude is most pronounced.

Table 6 compares the geocodable rate for pre- and post-treatment addresses in each group by address quality. The extremely low percentages in the first three rows are to be expected based on the properties that define a geocodable address (also keep in mind that we began with all 'bad' addresses for the experiment). The percentages reported in the fourth row reveal the main area in which the retrieval group far excels beyond the group of administratively-cleaned addresses. Not only did the retrieval effort return more *complete* addresses (including a street name and number, city, state and zip), but those complete addresses also happened to be very *good*. All but 14% were successfully translated into geographic coordinates.

Table 6. *Interactive Geocoding Results for White Center after Treatment*

Address Quality	Administrative		Retrieval	
	n	Geocoded	n	Geocoded
Intersection	18	0.0%	1	0.0%
No Street Name or Street Number	42	0.0%	18	0.0%
No City or ZIP	11	9.1%	2	0.0%
All Primary Address Elements Present	35	45.7%	96	86.5%
Total	106	16.0%	117	70.9%

PRELIMINARY FINDINGS FROM PROVIDENCE

As described above, we developed new training materials for the Providence interviewer training aimed at improving the precision and mappability of addresses collected in the mobility series. While Providence data collection is still underway at the time of this writing, we repeated the analysis reported above for the five sites and White Center using preliminary Providence data.[12] Our findings must be considered tentative; however, they do offer an indication of the efficacy of the new training efforts and present opportunities for further investigation.

In keeping with the mobility rates observed across all sites between waves 1 and 2, more than half (53%) of all Providence respondents had moved since November 2005; 42% of those respondents moved twice during that time, and 30% of the Providence sample moved at least three times since November 2005. We again began our analysis by taking inventory of the components present in each address. As shown in Table 7, Providence addresses (n=542) resembled those collected in San Antonio in that none identified only an intersection in the street address field, but the percentage of intersection-only addresses in other sites was also very small – ranging from .78% in Des Moines to 7.7% in White Center. The preliminary Providence data

[12] It is important to note that the households interviewed in the latest stage of data collection tend to be those that were difficult to locate – most often, because they have moved since the previous wave of the survey. By analyzing the mobility series data before these interviews have been completed we speculate that the omitted addresses may actually be of better quality among this highly-mobile group, based on our findings from the HLM models (addresses reported by 'mover' households were 1.7 times as likely to be geocodable than those reported by occupants of 'stayer' addresses).

reveal a small increase in the preponderance of complete addresses: 73% have all elements present versus 70% for the five sites. This is roughly equivalent to the rate observed in San Antonio (74%) and slightly below the rate in Des Moines (76%).

Table 7. *Initial Assessment of Data Quality for Providence*

Address Quality	Percentage
Intersection	0%
No Street Name or Street Number	20.1%
No City or ZIP	6.6%
All Primary Address Elements Present	73.3%

Note: n = 542.

We then repeated the interactive geocoding procedure using the preliminary Providence data (see Table 8). The success rate was again slightly higher than among the five sites (70% versus 68%), with a particularly strong showing among the complete addresses: over 90% were geocodable. Further research is necessary in order to isolate any causal impact of the additional training materials and interviewer coaching that were put in place for Providence, but the preliminary data suggest that these efforts yielded modest improvements in data quality. The absence of intersection-only addresses at the time of this writing is perhaps most promising. Nonetheless, it is clear that the additional training materials do not eliminate the need for post-interview data retrieval and cleaning.

Table 8. *Interactive Geocoding Results for Providence*

Address Quality	n	Geocoded
Intersection	0	0.0%
No Street Name or Street Number	107	0.0%
No City or ZIP	38	57.9%
All Primary Address Elements Present	397	90.4%
Total	542	70.3%

DISCUSSION

The findings from our data cleaning experiment for White Center suggest that while re-contacting respondents and using sophisticated search tools like Accurint is more than twice as expensive as cleaning address data using free online tools and locating information from previous waves of data collection, it is indisputably more effective. Our results corroborate the general consensus among survey methodologists that respondents do not routinely store distinct memories of dates and event details (Wu, Martin, and Long 2001). Successful collection of self-reported addresses depends on an interactive process between the interviewer and respondent involving top-down, sequential, and parallel probing techniques (Belli, Shay and Stafford 2001). The expectations for a 'complete' address must be made clear to both the interviewer and respondent. We learned this lesson upon discovering the higher percentage of intersection-only addresses recorded in White Center (7.7% versus an average of 1.6% in the preceding four sites). While these data are indeed geocodable at the city and county level, interviewers would have likely employed different strategies to elicit a precise street address had we communicated the strict criteria demanded by our analytical goals.

The efficacy of the phone retrieval efforts is particularly noteworthy given that over a year had passed between the data collection period in White Center (winter 2009) and the retrieval period (spring 2010), meaning that some respondents were being asked to recall moves that occurred four years ago. This provides some indication that the maximum parameters for the retrospective reporting of addresses can be extended farther back in time than when asking respondents about less memorable or habitual events (i.e. number of sick days, use of area services, etc.). The 'age' of a previous address did not prove to impose a significant impact on the overall propensity to be geocodable based on our HLM models. Yet the number of addresses *was* significant, suggesting that respondents who have moved frequently in a short period of time are likely to have more difficulty recalling complete addresses. Overall we find that there is indeed an opportunity to improve the quality of self-reported address data without making changes to question design or the initial survey administration. While the *Making Connections* sites are not nationally representative, our observations from the retrospective address series may

inform future neighborhood-based surveys that have an interest in mobility but must also consider cost and logistical constraints as well as comparability with other datasets.

The White Center findings also support the argument that a primary role of the interviewer is to help the respondent reconstruct memories of past events and apply targeted techniques to compensate for memory loss. We suspect that at least two other factors contributed to the differences in address quality across sites beyond those accounted for in the HLM models. First, we suspect that there is a correlation between neighborhood size and the ability to recall good quality addresses. For example, the small size of the *Making Connections* neighborhood in Providence (3.4 square miles versus a median size of 7 square miles for the five sites), combined with the preponderance of multi-unit buildings versus single family homes, yields closer-proximity movement within an area characterized by well-marked streets on a systematic urban grid. We hypothesize that it may be less taxing to recall 'good' addresses in these areas because street names and intersections are 'rehearsed' more frequently in urban contexts. Second, the absence of intersection-only addresses in San Antonio (the largest of the five sites, covering 24 square miles) may speak to residents' longer-term ties to the neighborhood. The median length of residence in their current neighborhood reported by San Antonio sample members was 10 years versus the five-site average of 6.4 years. We are interested in exploring further the possibility that respondents' tenure in the neighborhood or city may interact with neighborhood size and type of housing stock to impact respondents' ability to recall precise street addresses.

Future Research

A logical extension of the present research would be to reexamine the self-reported addresses collected in the wave 3 mobility series by employing a 'real-world' definition of data quality. One could extend the use of Accurint, as used by the retrieval team for our experiment with White Center data, or other administrative records to systematically validate respondents' presence at the addresses provided. Due to the challenges associated with establishing residence among highly-mobile populations with limited credit histories, this exercise has been less attractive to survey methodologists than multi-stage comparisons featuring reports of less ambiguous events like home purchases (for example, see Kennickell and Starr-McCluer 1997).

It is worth exploring whether the type of data retrieval that proved so successful in our experiment increases the risk of over-reporting addresses due to telescoping (Bradburn 2000) or other memory errors. Finally, additional questions in the mobility series gathered respondents' reasons for moving from their previous address and to their current address.[13] When combined with the geocoded addresses these data present rich opportunities to investigate the contextual factors shaping mobility and neighborhood change.

[13] Again, these were adopted from the MTO questionnaire.

REFERENCES

Belli, Robert; Shay, William; and Frank Stafford. Spring 2001. "Event History Calendars and Question List Surveys: A Direct Comparison of Interviewing Methods. *The Public Opinion Quarterly* 65(1): pp. 45-74.

Bradburn, Norman. 2000. "Temporal Representation and Event Dating." Included in *The Science of Self-Report*, pp. 49-61. Arthur Ston, Jaylan Turkkan, Christine Bachrach, Jared Jobe, Howard Kurtzman, and Virginia Cain, eds. Mahway, NJ: Erlbaum.

Fowler, Floyd. 1995. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, CA: Sage.

Kennickell, Arthur and Martha Starr-McCluer. October 1997. "Retrospective Reporting of Household Wealth: Evidence from the 1983-1989 Survey of Consumer Finances." *Journal of Business & Economic Statistics* 15(4): pp. 452-463.

Jobe, Jared B.; Tourangeau, Roger; and Albert F. Smith. 1993. "Contributions of Survey Research to the Understanding of Memory." *Applied Cognitive Psychology* 7(7): pp. 567-584.

Marcus, Alfred and Lori Crane. February 1986. "Telephone Surveys in Public Health Research." *Medical Care* 24(2): pp. 97-112.

Rabe-Hesketh, Sophia and Anders Skrondal. *Multilevel and Longitudinal Modeling Using Stata.* College Station, TX: Stata Press.

Schwenk, Charles. July 1985. "The Use of Participant Recollection in the Modeling of Organizational Decision Processes." *The Academy of Management Review* 10(3): pp. 496-503.

Suchman, Lucy, and Brigitte Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85(409): pp. 232-253.

Sudman, Seymour and Norman Bradburn. December 1983. "Effects of Time and Memory Factors on Response in Surveys." *Journal of the American Statistical Association* 68(344): pp. 805-815.

Wu, Lawrence; Martin, Steven; and Daniel Long. Summer 2001. "Comparing Data Quality of Fertility and First Sexual Intercourse Histories." *The Journal of Human Resources* 36(3): pp. 520-555.

APPENDICES

Appendix 1: Excerpt from Pre-Providence Training Materials Devoted to Mobility Series

How is this series of questions administered?

TRAINER: ASK FIs TO TAKE OUT THE BLANK MAIN INSTRUMENT AND TURN TO ITEM 1.9.

There are two main paths through this section: 1) a path for respondents who have not moved since Thanksgiving 2005 and 2) a path for respondents who have moved since Thanksgiving 2005.

Path for Respondent who has not moved since Thanksgiving 2005. This series begins with item 1.9. These respondents moved to their current address prior to Thanksgiving of 2005, so we do not gather any more addresses beyond this question. The interviewer question after item 1.9 will take us to item 1.12 (TRAINER: MAKE SURE FIs ARE FOLLOWING ALONG IN THEIR INSTRUMENTS).

At item 1.12, we will ask the respondents for the reasons they moved from their prior address to their current one. We then have two follow-up questions, items 1.14 and 1.15. Then, the interviewer instruction following item 1.15 will tell the interviewer that since the month and year at question 1.9 was prior to Thanksgiving of 2005 we skip to item 2.1.

Path for Respondent who has moved since thanksgiving 2005. Again, these respondents start with item 1.9, but this time, since the date on which they moved was after Thanksgiving 2005 these respondents will continue to item 1.10 where we will gather their prior address. These respondents will then follow the path through items 1.11, 1.12, 1.13, 1.15, and on to the two interviewer instructions.

These instructions look similar but they are different. At this point, the respondent has reported the dates of two moves. The first interviewer instruction checks to make sure that the first moving date given at item 1.9 isn't prior to Thanksgiving 2005. The second interviewer instruction checks to see if the moving date given at 1.11 is earlier than Thanksgiving 2005. If it is, the respondent skips to item 2.1 which is in the next section. If the date is after Thanksgiving 2005, the respondent continues with the mobility questions until they report a move that is prior to Thanksgiving 2005.

Let's try an example of each type of respondent.

TRAINER: LEAD ROUND ROBIN MINI-MOCKS. MOCK A WILL BE A RESPONDENT WITH NO MOVES. THIS IS A GREEN CASE THAT HAD NIR'S IN ROUNDS 1 AND 2. MOCK B WILL BE A RESPONDENT WHO HAS MOVED SINCE THANKSGIVING 2005. THE INHABITANTS AT THE ADDRESS JUST MOVED THERE IN AUGUST 2008 – AFTER OUR PRE-FIELD LOCATING EFFORT WAS COMPLETE.

Appendix 2: Training Guidelines for Recording Addresses

Normalizing Address, Phone Number and Email Data Entry

Why is this important?

- Normalization reduces errors in our mailing efforts and increases the likelihood of successfully reaching the respondent via the United States Postal Service
- Normalization eliminates the possibility of duplicate addresses in our records

ENTERING ADDRESSES:

What goes in Each Line of the Address?

Address Line 1:	The first line is for the number and street address (St / Ave / Blvd / Pkwy / Hwy). If the address is a PO Box, the PO Box and number would go on this line.
Address Line 2:	The second line is used in support of the first line. It can be used for company names, university names, care of (c/o), campus box numbers, other box numbers, or, if the address contains both a street address and a PO Box, the PO Box and number.
Apt:	This line is used for the unit type (Apt / Unit / # / Ste) followed by the number.
City:	This line is for the city, town or village name.
State:	This line is for the state name (see Job Aid: Abbreviations All in One").
Zip Code:	This line is for the zip code (5 digits).
Zip Code Ext:	This line is for the zip code extension (4 digits) if applicable.

QUICK ADDRESS RULES AND EXAMPLES:

Always ***use proper upper and lower case letters*** as described below:

- When writing out a Post Office Box, the first two words should be abbreviated into a capitalized "P" and "O," without periods, followed by the capitalized "Box," as such: "PO Box"
- Directionals should always be capitalized. When there are two letters in the abbreviation, they should both be capitalized without periods. For example, Northwest would be "NW" and South would be "S"
- Street suffix abbreviations usually begin with a capitalized letter, such as St, Ave and Blvd. A couple of exceptions to this rule for example are: RR for Rural Route or CR for County Road
- Use abbreviations instead of typing out the full street name such as for Street (St), Avenue (Ave), Boulevard (Blvd), Parkway (Pkwy), Highway (Hwy), Apartment (Apt), Suite (Ste). (See Job Aid: "Abbreviations All in One")

Here are examples of what your address entries should look like:

Street Addresses	Apartments	States
• 23 S Main St • 123 Park Ave NW • 123 Hwy 94 SW • 123 E 52nd St • 123 E North Ave (North is not abbreviated because it's the actual name of the street) • PO Box 1234	• Apt 58 L • Ste 1600 • Unit 708 • # 136 E • 1st Floor • Top Floor • Upper Level • Rear Apt	• IL, MI, CA, FL, GA, SC, NC • PR (Puerto Rico) • VI (Virgin Islands)

Appendix 3: Excerpt from "Spotlight on the Mobility Section" added for Providence Training

B. Probes and Job Aides

Probing is crucial to collecting high-quality data in the Mobility Series. A primary interest among researchers using the *Making Connections* data is mobility, or the movement of households across areas with different opportunities for economic and social advancement. This research depends on us collecting precise address information. If the respondent seems unsure of his or her answer or reports conflicting/incomplete information, please probe!

There are two components to the mobility section: 1) the date that they moved and 2) the address to which they moved. Let's discuss some effective probes for each one.

Date Probes

- For the question "In what month and year did you move to this address?" some example probes that could follow a "don't know" response include:
    - "Do you remember how many years ago you moved there?"
    - "How long did you live there for?" (e.g. "So you moved to the prior address in January of 2006. How long did you live at that address?")
    - "In what season of the year did you move?"
        - If they answer "summer", you can follow up with "Do you remember if it was June, July, or August?"
    - Ask what the weather was like when they moved. This may have made moving more difficult, so they may have strong memories of this.
    - "Did you move around any holidays or the beginning or end of a school year?"
- We have included a calendar in your bulk supply box. Please do not record any respondent information on the calendars, but do use this to help the respondent count the number of years or months they lived somewhere to determine when they moved to their next address.

Address Probes

- If the respondent is unsure of the address they were living at, we want to help them visualize the area around where they lived to help recall more specific information.
    - If they're unsure of the address, we have provided a map of our sample neighborhood. Ask if they can locate a school, park, or other landmark they lived near. Again, do not write any respondent information on the map, as this information is confidential.

    - If they live outside of the area of the map, ask if they can provide the name of a nearby school or church. From there, ask if they can remember the intersection where the building was at.
    - Ask if they can provide a range of addresses or block number (e.g. 3400 W Chester Ave.).
    - Clarify if they lived on a street, avenue, or something else if they don't provide that information.
    - Record the most accurate information you can. Use "X" to indicate that you probed for a given question and write a note next to the question to include the responses to your probes.
- Tell the R that you don't mind waiting while they look up old paperwork such as a bill, driver's license, etc. that may contain their previous address.
- Ask if a friend or other household member would be able to help remember the address.