Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications

Jennifer Sinibaldi

München 2014

Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications

Dissertation

Fakultät für Mathematik, Informatik und Statistik
Ludwig-Maximilians-Universität
München

Jennifer Sinibaldi

24 June 2014

First Chair: Prof. Dr. Frauke Kreuter

Second Chair: Dr. Gabriele B. Durrant

Defense date: 29 April 2014

Zusammenfassung

Repsonseraten in Haushaltssurveys, so ist weitläufig bekannt, sind weltweit im Fallen. Aufgrund von Nonresponse Bias in Surveys mit niedrigen Responseraten sorgt sich die Surveyforschung deshalb zu Recht um die Qualität der Statistiken, die auf diesen Umfragen basieren. Zwar versuchen einige Umfragen, etwa durch Responsive Design Techniken, den Nonresponse Bias schon während der Feldarbeit zu korrigieren, die meisten Surveys korrigieren jedoch im Nachhinein mit statistischen Anpassungen wie etwa Nonresponse Gewichtung. Um den Nonresponse Bias korrekt identifizieren zu können, müssen aber entweder während oder nach der Feldarbeit Daten zu Befragten und Nonrespondenten verfügbar sein. Entsprechende Daten sind nur beschränkt verfügbar, Surveyforscher nutzen deshalb häufig kommerzielle Marktforschungsdaten, administrative Daten und/oder Paradaten (Prozessdaten, die während der Feldarbeit gesammelt wurden).

Von diesen verschiedenen Datenquellen haben Paradaten den Vorteil, dass sie mit wenigen Kosten zu erheben sind, da sie entweder unbeabsichtigt während der Feldarbeit gesammelt werden oder sich sehr günstig für alle gesampelten Fälle erheben lassen. Zusätzlich lassen sich Paradaten während der Feldarbeit in beliebigem Rahmen sammeln, was dem Forscher die Flexibilität bietet, bewusst Variablen der Prozessdaten zu designen und zu sammeln, die mutmaßlich (erfolgreich) für Nonresponse Bias korrigieren können. Eine dieser flexiblen Formen von Paradaten zur Nonresponse Korrektur sind Beobachtungen von Interviewern. Diese, hauptsächlich in Face-to-face Studien gesammelten, Beobachtungen entstehen durch den Auftrag an die Interviewer, während der Feldarbeit Informationen zur Gegend, zum Haushalt oder zu ausgewählten Befragten festzuhalten. Diese Beobachtungen können, abhängig von den Nonresponse Korrelaten des jeweiligen Surveys, entsprechend angepasst werden.

Obwohl Beobachtungen von Interviewern augenscheinlich eine ideale Datenquelle zur Handhabung von Nonresponse Bias darstellen, können sie fehlerhaft sein. Da Messfehler in den Beobachtungen wahrscheinlich die Statistiken und Schlüsse, die von Nonresponse Korrekturen auf Basis dieser Beobachtungen gezogen werden, beeinflussen, sollten diese Fehler in den Beobachtungen der Interviewer nachvollzogen bzw. angesprochen werden. Die vorliegende Dissertation bemisst die Qualität von Interviewerbeobachtungen, indem die Eigenschaften der Messfehler in verschiedenen Nonresponse Kontexten untersucht werden. Die Arbeit ist in drei Abschnitte unterteilt. Jede Analyse baut auf der vorigen auf, wobei die Genauigkeit und Nützlichkeit von Interviewerbeobachtungen im Kontext von Nonresponse immer weiter untersucht werden.

In der *ersten Analyse* wird der Messfehler für fünf Interviewerbeobachtungen, die in Face-to-face Situationen gesammelt wurden, untersucht. Die Daten stammen aus der UK Census Nonresponse Link Study, die Interviewerbeobachtungen zu Haushaltseigenschaften, die im Rahmen von sechs UK Surveys gesammelt wurden, mit UK Zensusberichten zu den gleichen Haushalten verknüpft. Mit den Zensusdaten als wahren Werten kann die Genauigkeit der Interviewerbeobachtungen für Befragte und Nonrespondents der sechs Surveys gemessen werden. Zusätzlich werden Mehrebenenmodelle, die Haushalte, Gegend und Charakteristika der Interviewer berücksichtigen, genutzt, um die Korrelate der Genauigkeit zu untersuchen. Besondere Beachtung wird dabei der Reliabilität der Interviewer sowie Charakteristika der Interviewer, die die Genauigkeit beeinflussen, zuteil.

Die *zweite Analyse* nutzt die deutsche PASS Studie, um Interviewerbeobachtungen und kommerzielle Microm Daten mit dem Ziel zu bestimmen, welche der Daten genauer mit den wahren Werten des PASS übereinstimmen. Diese Analyse setzt sich weniger mit der Bestimmung des Messfehlers auseinander, sondern untersucht die Frage, ob die Interviewerbeobachtungen zentrale Befragungsergebnisse besser vorhersagen können als die kommerziellen Daten. Das signifikant wichtige an diesem Vergleich ist, dass hohe Korrelationen zwischen Hilfsvariablen und zentralen Befragungsergebnissen von hoher Wichtigkeit für effektive Nonresponse Anpassung sind. Die Analyse modelliert separat zwei Befragungsergebnisse, indem sowohl Interviewerbeobachtungen, als auch die Microm Daten als Prädiktoren verwendet werden und evaluiert die Signifikanz der Koeffizienten der beiden Datenquellen. Kreuzvalidierungen unterstützen die Ergebnisse der Modellierungen und lassen noch zuverlässigere Schlüsse zu, welche Datenquelle die Befragungsergebnisse besser vorhersagen kann.

Nach der sowohl direkt, als auch im Vergleich zu kommerziellen Daten abgeschätzten Genauigkeit mehrerer objektiver Interviewerbeobachtungen untersucht die *dritte Analyse*, inwieweit eine subjektive Interviewerbeobachtung zukünftiges Antwortverhalten von Personen im Survey Sample vorhersagen kann. Dabei ist die bei jedem Kontakt einer Telefonumfrage getroffene Einschätzung des Interviewers, ob eine gegebene Person am Survey teilnehmen wird, die untersuchte Interviewerbeobachtung. Die Analyse vergleicht den Fit und die Diskriminierung „klassischer“ Response Propensity Modelle, die alle Anrufdaten und Interviewer Charakteristika einschließen, mit Propensity Modellen mit Interviewerbeobachtungen und bestimmt, ob diese Beobachtungen die Propensity Modelle signifikant verbessern können. Die Performance der Interviewerbeobachtung wird in einem simulierten Responsive Design getestet, wo Propensity Modelle täglich generiert werden, um zu bestimmen, ob die Beobachtung die Genauigkeit dieser täglichen Vorhersagen gegenüber „klassischen“ Propensity Modellen signifikant verbessern kann. Alle Analysen nutzen zeitdiskrete Hazard Modelle und kontrollieren für Zufallseffekte der Interviewer.

Die Ergebnisse dieser Analysen stellen ein dringend benötigtes Benchmark für die Qualität von Interviewerbeobachtungen dar. Mit den Informationen zu den Korrelaten von Messfehlern der Beobachtungen können Anwender von Surveys zusätzlich Fortschritte bei der Verbesserung der Qualität von Interviewerbeobachtungen machen. Die Ergebnisse untermauern zudem die Effektivität von Interviewerbeobachtungen bezüglich der Anwendung in Nonresponse-Szenarien sowohl während, als auch nach der Datensammlung und deuten auf das Potential von Interviewerbeobachtungen für diese Zwecke hin, sobald die Qualität verbessert ist.

Abstract

It is widely known that response rates to household surveys are falling throughout the world. In this environment, survey researchers are concerned about the quality of the statistics generated from surveys with low response rates due to the presence of nonresponse bias. Although some surveys make efforts to correct for nonresponse bias during survey fieldwork (e.g., through responsive design techniques), most surveys correct the data after fieldwork closes, using statistical adjustment (i.e., applying nonresponse weights). To address nonresponse bias, either during or after fieldwork, data must be available for both the respondents and nonrespondents. The options for data with this property are limited but survey researchers generally turn to commercial marketing data, administrative records, and/or paradata (the process data collected during survey fieldwork).

Of these three forms of auxiliary data, paradata have the appealing benefit of being low cost since they are either unintentionally generated during survey fieldwork or very inexpensive to collect for all sampled cases. In addition, the range of paradata that can be captured during survey fieldwork is not fixed, giving researchers the flexibility to intentionally design and collect process data variables that are expected to successfully correct for nonresponse bias. One of these flexible forms of paradata used in nonresponse applications is interviewer observations. Most often collected in face-to-face studies, interviewer observations are generated by asking interviewers to record characteristics about the area, household, or selected respondent when they visit a property. These observations can be customized to the survey, depending on the correlates of nonresponse for that survey.

Although a seemingly ideal data source for addressing nonresponse bias, interviewer observations may be inaccurate. Since measurement error in the observations is likely to affect the statistics and conclusions drawn from nonresponse applications using the observations, the error properties of interviewer observations should be understood, if not addressed. This dissertation assesses the quality of interviewer observations by examining their measurement error properties in various nonresponse contexts. Presented *in three parts*, each subsequent analysis builds on the previous by further exploring the accuracy and utility of interviewer observations for nonresponse applications.

The *first analysis* investigates the measurement error of five interviewer observations commonly collected in face-to-face data collections. The data come from the UK Census Nonresponse Link Study, which links interviewer observations of household characteristics reported for six UK surveys to UK Census reports for the same households. Using the Census data as the true value, the accuracy of the interviewer observations can be assessed for both respondents and nonrespondents to the six surveys. In addition, multilevel modeling, incorporating household, area, and interviewer characteristics, is used to explore the correlates of accuracy. Special attention is given to the reliability of interviewers and the characteristics of interviewers that affect accuracy.

The *second analysis* uses the German PASS study to compare interviewer observations to Microm commercial data to determine which shares more (accurate) information with the true values reported in the survey data. A kind of measurement error assessment, this analysis is not concerned with directly evaluating the level of error but instead, determining if the interviewer observations are more predictive of key survey outcomes than the commercial

data. The significance of this comparison is that high correlations between auxiliary data and key survey outcomes are important for effective nonresponse adjustment. The analysis separately models two survey outcomes, using both interviewer observations and Microm data as predictors, and evaluates the significance of the coefficients from these data sources. Cross validation is used to support the findings from the modeling and more confidently conclude which source is more predictive of the survey outcomes.

Having assessed the accuracy of several objective interviewer observations both directly and relative to commercial data, the *third analysis* investigates a subjective interviewer observation for its ability to accurately predict the future response behavior of people selected into the survey sample. The interviewer observation explored is the interviewer's assessment of the likelihood that a given respondent will participate in the survey, collected at each contact of a telephone study. The analysis compares the fit and discrimination of "classic" response propensity models, which include call record data and interviewer characteristics, to propensity models including the interviewer observation, to determine whether these observations significantly improve the propensity models. The performance of the interviewer observation is tested in a simulated responsive design, where propensity models are generated daily, to determine if the observations significantly improve the accuracy of these daily predictions over using "classic" propensity models. All analyses use discrete time hazard models, controlling for the random effect of interviewers.

The findings from these analyses will provide a much needed benchmark for the quality of interviewer observations. In addition, with information on the correlates of measurement error of the observations, survey practitioners can take steps to improve the quality of interviewer observations. The results also support the effectiveness of interviewer observations in nonresponse applications both during and after data collection and indicate the future potential of interviewer observations for these purposes once the quality is improved.

Table of Contents

Chapter 1: Introduction

1.1 Understanding nonresponse bias

It is well-known that response rates to household surveys have been declining (de Leeuw and de Heer 2002; Curtin et al. 2005). Although a higher response rate does not guarantee an improvement in nonresponse bias (Keeter et al. 2000, 2006; Groves and Peytcheva 2008), capturing less of the sampled population generally heightens concerns of nonresponse bias in the survey statistics. Nonresponse bias will be present for some statistics if the nonrespondents have different values for the survey variables of interest than those given by the respondents. This is represented by the deterministic nonresponse bias equation (Groves and Couper 1998):

$$Bias(\bar{y}_r) = \left(\frac{M}{N}\right)(\bar{Y}_r - \bar{Y}_m)$$

where

$\bar{y}_r$ is the mean for a particular characteristic *y* among the respondents to the survey

$\frac{M}{N}$ represents the proportion of nonrespondents to the survey

$\bar{Y}_r$ is the mean of the particular characteristic in the population among the respondents

$\bar{Y}_m$ is the mean of the particular characteristic in the population among the nonrespondents.

Although the above equation is helpful to understand when nonresponse bias is present, especially after all data have been collected and individuals can be easily grouped into respondents and nonrespondents, the concept of sampled individuals being pre-determined as definitely a respondent or nonrespondent is overly simplistic. Instead, researchers prefer to characterize sampled individuals in terms of likelihood, or propensity, to respond which allows for a better than zero chance of convincing every sampled individual to cooperate. Embracing this theory, researchers tend to favor the stochastic model of nonresponse bias (Bethlehem 1988) which estimates bias using a probability to respond rather than a deterministic indicator.

$$Bias(\bar{y}_r) \approx \frac{\sigma_{yp}}{\bar{p}}$$

$\bar{y}_r$ is the mean for a particular characteristic *y* among the respondents to the survey

σ_{yp} is the covariance between the particular characteristic in the population and the response propensity

$\bar{p}$ is the mean of the response propensities in the population

This expression relates the magnitude of the bias to the correlation between the characteristic of interest and propensity to respond. If this correlation is weak because people's decisions to respond are not related to the characteristic, the bias will be small. In more extreme circumstances, if there is no relationship between *y* and *p* (the correlation =0), or the values of either *y* or *p* are the same for all individuals in the population (the variance = 0), nonresponse bias is nonexistent. This relationship between *y* and *p* is important in nonresponse applications used both *during* and *after* data collection (as will be explained below).

Theories of response and nonresponse bias are still evolving, however. The equations presented above are intended to characterize the propensity to respond as a fixed characteristic of the individual under specific survey conditions. More recently researchers have begun to characterize the likelihood to respond as a dynamic process (Olson and Groves 2012) where an individual's likelihood to respond fluctuates throughout the field period, depending on changes to survey protocol or treatment. The stochastic equation above is typically used to capture this change by re-estimating the propensities for each individual as fieldwork progresses. This technique and other nonresponse applications are presented in the next section.

1.2 Nonresponse applications

Efforts to correct for nonresponse bias can be made both during and after data collection. During data collection, popular techniques to improve the representativeness of the respondent pool (and thereby reduce the nonresponse bias) are responsive survey design and adaptive survey design. Both of these techniques involve introducing changes to survey protocol in order to improve the quality of the statistics but with cost efficiency in mind. Although the choice of terminology can sometimes be a matter of preference by a researcher or organization, the general distinction between adaptive and responsive survey designs seems to be when the protocol change is applied. Adaptive designs tend to learn from prior waves and adapt the protocol for the next wave while responsive designs closely monitor various indicators during the data collection and introduce protocol change(s) at designated points during fieldwork (Schouten 2013).

The idea to deliberately and intelligently adjust protocol during data collection was first outlined in an article by Groves and Heeringa (2006). Their responsive survey design technique divided the survey field period into design phases where at the end of each phase, a protocol change was made. The use of experimentation in the early phases is encouraged as well as monitoring of key survey statistics, and effort and productivity indicators. The surveys on which they implemented responsive design were all face-to-face.

Since that groundbreaking article, other researchers have documented strategies executed in the spirit of responsive design, adjusting the technique to suit their organizational or survey needs. Peytchev et al. (2009) introduced protocol changes in the form of incentives and a shorter questionnaire to improve nonresponse bias on a telephone survey. Wagner (2013) used timely call record data to continuously predict the best time to contact a case in both a telephone and face-to-face survey. Laflamme et al. (2008) "actively managed" the survey data collection of two telephone surveys to immediately respond to indicators of low quality or high cost.

Regardless of whether survey managers chose to implement a responsive or adaptive design or not *during* the data collection, at the close of data collection efforts to correct for potential nonresponse bias are often performed in the form of nonresponse weighting. Post-survey adjustments can be developed based on characteristics of the cases selected into the probability-based sample if auxiliary data are available for each case. One way to perform this type of sample-based weighting is to use propensity modeling. Devising the weights involves estimating the propensity to respond for each case in the sample using a logistic regression model predicting cooperation. This is often done in two parts, corresponding to the work of Groves and Couper (1998) (discussed later), estimating first the propensity of

contact and then, conditional on contact, the propensity of cooperation. Either using the raw propensities or grouping them into adjustment cells, the inverse of the propensity to respond is applied to each respondent to adjust for the nonresponse bias (Little and Rubin 2002; Kalton and Flores-Cervantes 2003). The sources of data that can be used to calculate sample-based nonresponse weights are discussed in the next section.

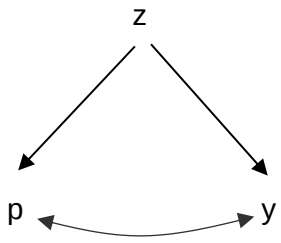
1.3 Characteristics of data used in nonresponse applications

For both responsive design and sample-based post-survey adjustment, data must be available for every eligible unit in the sample frame, both respondents to the survey and nonrespondents. Finding such data can be challenging. Options include administrative record data which can be linked to the sampling frame if available for the entire population. Also, commercial marketing data available from companies such as Experian in the United States or microm in Germany, may offer characteristics at the neighborhood or household level that can be appended to the sample frame for each case. An additional option is survey process data, called paradata (Couper 1998). Paradata are generated during survey data collection and therefore, are not part of the sampling frame but available after data collection begins. Examples of paradata include: times and dates of call attempts made to the household, records of keystrokes made by the interviewer or respondent, or pre-categorized notes by the interviewer recording respondent concerns or comments about the survey request. The paradata available vary by mode of data collection with some data being available only for a particular mode (e.g., mouse movements in a web survey, interviewer observations of area characteristics in a face-to-face survey). See Kreuter 2013 for details of different types of paradata currently available.

The type of data collected and used in nonresponse applications are motivated by theories of the mechanisms of nonresponse. Much of this research has been guided by the work of Groves and Couper (1998), who emphasize two distinct stages of the response process: making contact and then, conditional on making contact, gaining cooperation. Making contact is a function of first, the at-home patterns of the householder and second, impediments to accessing the householder such physical barriers of entry to the property in face-to-face data collections or technology inhibiting telephone contact (e.g. caller ID) in a telephone data collection. Data that may help characterize at-home patterns are the days of the week and time of day of the contact attempts along with the corresponding outcome of the attempt, as well as interviewer observations of lifestyle like presence of children and the absence of cars during working hours. Regarding access impediments, interviewers can observe the type of building and the presence of entrance intercoms and locked gates or commercial data can indicate whether a telephone number is publicly listed or not. Once contact is made, gaining cooperation is a function of the social environment, survey design features, characteristics of the householder and interviewer, and the interaction between the respondent and the interviewer. Data that may help characterize the social environment are population density and crime statistics in the area. Important survey design features are the survey topic, length of data collection, and offering an incentive or not. Characteristics of the interviewer can be collected using an interviewer survey or gleaned from personnel records while characteristics of the householder such as age, race, and income can be captured using interviewer observations or commercial data. The interviewer-respondent interaction is most often captured through contact observations recorded by the interviewer.

As detailed above, there are many options for collecting data that could potentially capture characteristics of known correlates of nonresponse and therefore be useful for estimating

response propensity or for responsive survey design applications. However, in order to be effective for nonresponse adjustment, the data must be correlated with the propensity to respond and the key variable(s) of interest (Little and Vartivarian 2005). If the key survey estimate is *y*, the propensity to respond is *p*, and a variable used for adjustment is z, the relationship between these three variables can be represented as:



This is called the Common Cause Model (Groves 2006) and although helpful for thinking about nonresponse adjustment, finding *z* variables with strong correlations with both *y* and p is challenging. Often variables are highly correlated with either *y* or *p* but not both (Kreuter et al. 2010b; Kreuter and Olson 2011).

1.4 Paradata for nonresponse applications

This difficulty finding variables highly correlated with response and the survey outcome(s) has led researchers to more closely investigate paradata. Although pardata were originally touted as a tool for measuring and monitoring the quality of the data collection process and the resulting survey data (e.g., using time stamps for indications that the interviewers are administering the questionnaire too quickly or falsifying surveys; see Couper 1998), these data have shown potential for use in nonresponse adjustment models (Beaumont 2005) as well as responsive survey designs (Kirgis and Lepkowski 2013).

Paradata are attractive options for nonresponse applications for several reasons. First, there is minimal or *no cost* to collect or capture these data – for the most part, these data are automatically generated during the survey process. Second, if technological systems are designed to process the paradata as they are generated, paradata can also be *timely* providing the necessary up-to-date information needed for data collection decisions. Lastly, researchers have a level of *control* over the structure and quality of paradata that is appealing. Researchers can not only choose what variables and indicators to collect but the level of detail at which they are collected[1]. By using technology to control the input of the paradata, some effort can also be made to control the quality. In situations where the data are not of acceptable quality, the researcher is often aware of the flaws and the reasons for them. All of these features give paradata an advantage of commercial data, which must be purchased, can be out-of-date, and allows no control over the quality of its collection. In addition, researchers must apply their own efforts to assess the quality of these data and their fit for the purpose (see the introduction to chapter 3 in this document for details on the quality limitations of commercial data). When the commercial data are found to be lacking in quality, the researcher has limited or no ability to improve them.

[1] For example, contact observations can be collected in detail such as in the CHI (see Bates et al. 2010) or aggregated to broader categories such as “positive statements”, “negative statements” and “questions”.

1.5 A specific form of paradata: Interviewer Observations

One of the more versatile forms of paradata, in terms of allowing the researcher to design and somewhat control the quality, are observations of area, household, or person characteristics recorded by interviewers; so-called *interviewer observations*. Since an interviewer is necessary to gather this type of data, interviewer observations are only available in the interviewer administered modes-- face-to-face and telephone data collections.

As detailed above, interviewer observations can be collected to capture characteristics correlated with response. Their flexibility also allows for the design of observations that closely, if not exactly, match the survey variables of interest. This is the unique advantage of interviewer observations over other forms of paradata and researchers are currently designing and collecting observations to match variables of interest in their surveys. One example of a survey with observations designed to be highly correlated with *y* is the National Survey of Family Growth, a study of fertility, family formation and risks of sexually transmitted disease in the United States. In this face-to-face study, interviewers are asked to observe evidence of children in the household and guess as to whether the selected respondent is in a sexually active relationship (West 2013a). In the Los Angeles Family and Neighborhood Survey (LAFANS), a study of the quality of life in LA neighborhoods, interviewers made neighborhood observations about the evidence of graffiti and trash (Casas-Cordero et al. 2013). As discussed in chapter 3, the Panel Study of Labor Market and Social Security (PASS) in Germany collected observations corresponding with two of the key outcomes of this study: whether the household was on unemployment benefits (UB2) or not and the general income level of the household.

Although the carefully designed interviewer observations may seem to be an ideal solution to address nonresponse bias, the reality is that this type of paradata (as well as other types) suffers from measurement error. Analyses from Fuller (1987) and Carroll et al. (2006) show that measurement error in the variables used in the derivation of survey statistics can introduce bias in those statistics. In the context of this body of work, the statistics of concern are the predicted probabilities of response and the nonresponse weights calculated from those probabilities. Therefore, errors in the interviewer observations may undermine the ability of these paradata to correct the nonresponse bias. Considering this, the quality of the observations and the resulting impact on the statistics should be evaluated.

Although accuracy of interviewer observations has been studied sporadically over the years (e.g. Campanelli et al.1997, Pickering et al. 2003) the deliberate study of measurement error in interviewer observations (and paradata in general) is relatively new, especially with an emphasis on studying the effect of this error on nonresponse applications. Much of the recent knowledge in this area has resulted from the work of Brady West and colleagues (e.g. West 2013a, West et al. *forthcoming*) as well as the analyses presented here, in this dissertation (Sinibaldi et al. 2013; Sinibaldi et al. *forthcoming*). An additional significant contribution in the study of error in interviewer observations taken over the phone has come from McCulloch et al. (2010). (For a complete review of literature on the quality of interviewer observations and other forms of paradata, see West and Sinibaldi 2013).

But assessing the measurement error is the just the first step in understanding the value of interviewer observations for nonresponse applications. Besides identifying the presence and magnitude of measurement error, researchers must understand the correlates of the error to devise solutions to improve the quality of the observations. The field would also benefit from knowing how the magnitude of error in the observations compares to the magnitude of error in other data that could be or is used for the same purpose. Lastly, the impact of the error on the intended application (e.g. estimates of key survey statistics, classification of the probability of a case to respond) will help prioritize the options for dealing with the error (e.g., taking steps to improve the quality of the observation, tolerating the error as it is, or abandoning the use of the observation all together). This is a sizeable research agenda.

All of these questions pertaining to better understanding the measurement error of interviewer observations have motivated the work presented here. This dissertation "Evaluating the Quality of Interviewer Observed Paradata for Nonresponse Applications", follows the research agenda above, taking each question as a separate analysis but acknowledging that there is a necessary order to the research questions. The findings from the analyses designed and conducted in order to answer the first question, lay the foundation of knowledge necessary to address the second question, and so on for the third question. Together, the three papers provide a coherent body of work on the quality and utility of interviewer observations for nonresponse applications.

1.6 Outline of the three analyses

Paper 1, Chapter 2
In the first analysis, I use a dataset from the United Kingdom that links interviewer observations taken on six different surveys to the 2001 Census data for the households selected for those surveys. This linkage provides a gold standard for the observations, allowing me to evaluate the accuracy of five typically recorded interviewer observations for both respondents and nonrespondents. Using multilevel modeling to account for the clustering of households within interviewers and areas, I examine correlates of the accuracy of the observations using characteristics of the area, household, and interviewer. The results find that the measurement error of the interviewer observations is minimal (i.e. the accuracy is high) and correlates of accuracy pertain to the visibility of the property and the level of interviewer-respondent interaction. This paper has been published in a special issue of Public Opinion Quarterly on measurement error (Sinibaldi et al. 2013).

Paper 2, Chapter 3
In the second analysis, I compare the performance of two data sources commonly used for nonresponse adjustment, interviewer observations and commercial data, to determine which is the better option for this purpose given that both suffer from measurement error. The analysis uses German PASS data to evaluate the ability of these data types to predict two key survey outcomes: whether someone in the household is on UB2 and the level of household income. Being a better predictor would correspond to the data being a better candidate for nonresponse adjustment (assuming the correlations with response are similar between the two data sources). The analysis finds that interviewer observations are better at predicting these outcomes, particularly for the special subpopulation that this survey targets. This paper has been accepted for publication in the summer 2014 issue of Public Opinion Quarterly.

Paper 3, Chapter 4
In the third analysis, I examine a new call-level interviewer observation for its ability to significantly improve the predictive power of response propensity models, particularly those used for directing fieldwork when conducting responsive survey design. This analysis addresses the final research question on the agenda concerning the impact of the error on the intended application (i.e. responsive survey design). The interviewer observation is subjective and therefore especially prone to measurement error but it is used to predict an objective outcome, cooperation at the next contact. I find that the new observation does improve the propensity models and seems to better predict cooperation at the next contact than the forms of paradata typically used, especially at the end of the field period. However, the predictions are not perfect and if the error in the observation was reduced, it could further improve the predictive power of the models.

Combined, these papers take a first cut at addressing the research agenda outlined above. All of these analyses investigate the quality of interviewer observations in different contexts, report conclusive findings, and produce practical recommendations. Although there will be further research in each of these areas, this body of work will (and has already) provided the field with a much needed benchmark for the quality and utility of interviewer observations. In the conclusion of this dissertation, I summarize the findings and lessons learned across the three analyses and propose next steps for furthering this course of research, presenting a revised research agenda for the future.

Chapter 2: Evaluating the Measurement Error of Interviewer Observations

Note: This is a pre-copy-editing, author-produced PDF of an article accepted for publication in *Public Opinion Quarterly* following peer review. The definitive publisher-authenticated version

> Sinibaldi, Jennifer, Gabriele B. Durrant, and Frauke Kreuter. 2013. "Evaluating the Measurement Error of Interviewer Observed Paradata." *Public Opinion Quarterly*, 77: 173-93. doi: 10.1093/poq/nfs062

is available online at: http://poq.oxfordjournals.org/content/77/S1/173.full.

2.1 Introduction

Faced with high nonresponse rates, survey researchers use nonresponse adjustment methods (Kalton and Flores-Cervantes 2003) and, more recently, responsive survey designs (Groves and Heeringa 2006) to address potential nonresponse bias. For both techniques, paradata (Couper 1998; see Kreuter and Casas-Cordero [2010] for a detailed review) are now used (e.g., Lepkowski et al. 2010) or are being explored (e.g., O'Hare 2012). In face-to-face surveys, interviewers can observe housing unit and neighborhood characteristics for both respondents and nonrespondents, making these observations potentially good candidates for the identification and correction of nonresponse bias. Major surveys in the United States (e.g., National Survey of Family Growth [NSFG]; Lepkowski et al. 2010), United Kingdom (e.g., Understanding Society; McFall 2011), and mainland Europe (e.g., European Social Survey [ESS]; Jowell et al. 2007), now ask interviewers to make observations on a routine basis.

However, using interviewer observations successfully depends on their quality. Analyses from Fuller (1987) and Carroll et al. (2006) document the impact on final statistics when variables containing measurement error are used in the derivation of survey estimates and statistical modeling. In a study particular to paradata, Biemer et al. (2013) found that slight inaccuracies in the number of calls reported could produce significant bias in predicted response propensities. Yet research specifically on the measurement error of interviewer observations is scarce. There are some studies that assess the reliability of interviewer observations (e.g., Casas-Cordero et al. 2013; Sinibaldi 2010; Weich et al. 2001), but studies on their validity are rarer still. West (2010, 2013a) examined measurement error in the observations of NSFG interviewers by comparing them to survey and household roster data. However, validation of the observations was not possible for uncooperative cases. In the absence of a gold standard, Bates et al. (2010) used alternate indicators of accuracy, such as the time elapsed between contact and the entry of the observations, for their assessment of measurement error. Pickering et al. (2003) investigated the accuracy of the interviewer observations for all sample units but did not model possible correlates of accuracy. Thus, there is still a need to examine the accuracy of interviewer observations, for both respondents and nonrespondents, and explore possible determinants of measurement error.

In the data, interviewer observations for both respondents and nonrespondents are linked to UK Census data. Therefore, the accuracy of the interviewer observations can be assessed by analyzing each observation's agreement with a criterion that is assumed to be accurate: respondents' self-reports of the same characteristics from the UK Census. Furthermore,

housing unit, neighborhood, and interviewer characteristics linked to the data provide potential predictors of inaccuracies.

Although what interviewers are asked to observe can vary from survey to survey, the accuracy of a set of routinely collected observations is reported. If there is minimal error in the observations, there is little concern about using them to correct for nonresponse bias, assuming they predict survey participation and substantive response (Little and Vartivarian 2005). However, if measurement error is high and the variables seem to be related to nonresponse bias, then survey researchers need to identify ways to improve the collection of these data. The identification of common covariates that affect the accuracy of observations can inform these improvement efforts.

2.2 Data

In the UK Census Nonresponse Link Study, data from the 2001 UK Census are linked to paradata from six major, face-to-face, UK household surveys collected at approximately the same time as the Census (April 2001). For the households selected for each of the six surveys, the linked paradata include call record information and observations interviewers made during data collection about housing unit, neighborhood, and household characteristics. Also linked to the data were the following: aggregate area-level Census information, interviewer characteristics, and interviewers' reports of their attitudes and the approaches they used at the doorstep. Interviewers' reports were taken from a survey of all interviewers working for the survey agency, the UK Office for National Statistics, which was conducted at the time of the 2001 Census (Freeth et al. 2002). For additional details about the dataset, see Beerten and Freeth (2004) and Durrant and Steele (2009).

The wording of several questions on the observation forms closely corresponded to the wording for the Census questions. Thus, respondents' self-reports from the Census can be compared to the interviewer observations to assess the accuracy of the observations for both respondents and nonrespondents to the surveys. These observations are about (1) the physical structure of the housing unit (Type of HU); (2) whether the unit is public housing, also known as a council property (Council); (3) the employment status of at least one adult (Working); (4) the race/ethnicity (White) of the household; and (5) the presence of at least one child (Children). Table 2.1 provides the original questions from the interviewer observation form and the Census questionnaire, as well as the recoding of the response options for the analysis. Of these observations, two—Type of HU and Council—pertained to the housing unit and did not require contact with the household. Because the other three observations required contact, interviewers were asked to record the characteristic only for refusing and cooperative cases. Interviewers did not receive specific training about how to make the observations.

Respondents' self-reports are assumed to be a reasonably accurate reflection of the true values, because householders would be able to correctly report the information (e.g., the type of accommodation they live in), and the timing of the survey data collection relative to the Census data collection limits the possibility of true change (e.g., the birth of a first child between participation in the Census and the interviewer observation). However, there is the possibility of measurement error in the Census self-reports. For example, the respondent may misreport his/her employment status due to misunderstanding or ignoring the details of the Census definition, or fear of reporting illegal work. However, since only one adult in the

household has to be working in order to match an interviewer's observation of any adult in paid work, this reduces the impact of any possible measurement error in the Census information.

In some situations, the observations and Census questions may not ask for exactly the same information. For example, one would expect a house on a council estate, as the observation asks, to be rented from the local authority. In fact, some individual units within a council estate may be owned outright. In addition, there may be individual units outside the established estates that are used for public housing (Department for Communities and Local Government 2012b). Therefore, an interviewer may make an accurate observation that a property is on a council estate, but the Census information may indicate that it is privately owned. Another example of a potential mismatch may result from slightly different definitions of "dependent child." On the Census form, a dependent child is aged 0 to 15 or aged 16 to 18 in full-time education (Office for National Statistics 2004). In the interviewer observations, interviewers are asked to record only the number of children under 16 years old. These discrepancies will be considered when evaluating the results.

The analytic sample comprises all households selected for interviewing during May–June 2001, the months immediately following the Census that were successfully linked to the paradata (95-percent linkage rate). Certain cases, such as noneligible persons and vacant homes, were deleted from the dataset (see Durrant and Steele 2009). The observations that were collected only for contacted cases did not include noncontacts[2] by definition. For the models of accuracy, the analysis focuses on correct and incorrect observations only, and drops the missing cases (see appendix 2A for elaboration). Overall, the missing-data rates were low, with few missing observations for Type of HU (0.3 percent) and White (1.5 percent) (see table A1, appendix 2A). The Children observation had the most missing data (12.0 percent). Finally, to provide enough data for the estimation of interviewer variance, interviewers who provided observations for fewer than three cases were removed from the analysis. Depending on the observation, the loss ranged from five to twenty-three interviewers,[3] resulting in a case base of approximately 15,000 to 18,000 households for the five interviewer observations (see table B1 in appendix 2B).

Each household in the dataset is attributed to one interviewer and one area, defined as the area governed by the local authority.[4] For the Council observation, eight areas that had fewer than three cases were collapsed with the bordering area with the fewest households to provide a minimum sample size for variance estimation. Depending on the number of cases dropped for the analysis of each observation and the collapsing of areas, there are 537 to 560 interviewers and 384 to 392 areas. Since this dataset combines several surveys,

[2] Noncontact is a result code assigned by the interviewer, but it does not necessarily mean that there was absolutely no contact with the household during fieldwork. A case that does not refuse but avoids further contact is often coded as a noncontact. Also, if contact is made with the household but never with the selected respondent, the case is coded as noncontact. Therefore, the effect of the noncontacts in this analysis may be underestimated.

[3] The loss was five interviewers for the Type of HU observation, seven interviewers for the Working and White observations, nine interviewers for the Council observation, and twenty-three interviewers for the Children observation.

[4] The local level of government in the UK is called a local authority district. Examples are Oxford (non metropolitan district), Liverpool (metropolitan district), Southampton (unitary authority), and Camden (London borough) (Department for Communities and Local Government 2012a). The local authority, also called a council, provides the public or "council" housing, as in the observation analyzed in this study (Department for Communities and Local Government 2012b).

Table 2.1. Wording and Response Categories for Interviewer Observations and Respondents' Census Reports, Showing Recoding Used in the Analysis

Response options used in analysis	Question on interviewer observation form	Question on Census form
Type of HU	**What type of accommodation is it?**	**What type of accommodation does your household occupy?**
1 House	**House or bungalow** Detached Semi-detached Terrrace/end of terrace	**A whole house or bungalow that is:** Detached Semi-detached Terraced (including end-terrace)
2 Flat	**Flat or maisonette** In a purpose-built block Part of a converted house/some other kind of building Room or rooms	**A flat, maisonette, or apartment that is:** In a purpose-built block of flats or tenement Part of a converted or shared house (includes bed-sits) In a commercial building (for example, in an office building or hotel or over a shop)
3 Caravan, other	Caravan, mobile home, or houseboat	**Mobile or temporary structure:** A caravan or other mobile or temporary structure
Council	**Is the house/ flat part of a council or housing association housing estate?**	**Who is your landlord?**
1 Council house	Yes, part of a large council estate Yes, part of a council block	Council (local authority) Housing Association, Housing Cooperative, Charitable Trust, Registered Social Landlord
0 Not	No	Private landlord or letting agency Employer of a household member Relative or friend of a household member Other

Table 2.1. *Continued*

Response options used in analysis	Question on interviewer observation form	Question on Census form
Working Adult	**Is any adult in paid work?**	**Number of adults in employed work** (recoded by Office for National Statistics)
1 Working adult	Yes	1 or more
0 Not	No	0
White	**Do you know or think the occupants are:** (Code from observation; code ALL that apply)	**What is your ethnic group?** (Choose ONE)
1 All White	White	White (British, Irish, any other White background)
0 Not	Mixed	Mixed (White and Black Caribbean, White and Black African, White and Asian, any other Mixed background)
	Asian (Indian, Pakistani, Bangladeshi)	Asian or Asian British (Indian, Pakistani, Bangladeshi, any other Asian background)
	Black (Caribbean, African, other)	Black or Black British (Caribbean, African, any other Black background)
	Chinese or other ethnic group	Chinese or other ethnic group (Chinese, any other)
Children	**Number of children (less than 16 years)** (open-ended question)	**Number of dependent children** (recoded by Office for National Statistics)
1 Child present	1 or more	1 or more
0 Not	0	0

interviewers are generally not exclusive to one area. This produces a cross-classified nested structure of interviewers and areas and avoids a complete confounding of area and interviewer effects (see Durrant et al. 2010). (See figure C1 in appendix 2C for an illustration of the nesting pattern of interviewers and areas.)

2.3 Methods

The validity of the observations is evaluated by analyzing the agreement between the interviewer observations and their equivalent records from the Census, using multilevel cross-classified logistic regression models. Such models are necessary to account for the clustering of households within interviewers and areas. Standard logistic regression would lead to underestimation of standard errors and therefore incorrect inference, especially for higher-level variables (here, interviewer and area characteristics). Furthermore, multilevel models allow the investigation of substantive research questions, such as the reliability of the observation, by producing statistics about how much variation is due to interviewers (Schnell and Kreuter 2005).

Let $y_{i(jk)}$ denote agreement between a particular observation and the criterion (i.e., "without measurement error") for household i contacted by interviewer j in area k, where the parentheses indicate the cross-classification of interviewers with areas. The dependent variable is coded as

$$y_{i(jk)} = \begin{cases} 1 & \text{without measurement error} \\ 0 & \text{with measurement error} \end{cases}$$

The multilevel cross-classified logistic model for the occurrence of measurement error (taking "with measurement error" as the reference category) can be written as

$$\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) = \beta^T x_{i(jk)} + u_j + v_k \qquad (1)$$

where $\pi_{i(jk)}$ denotes the probability of accurate measurement, $\pi_{i(jk)} = \Pr(y_{i(jk)} = 1)$; $x_{i(jk)}$ is a vector of household, interviewer, and area-level covariates; β is a vector of coefficients; and u_j and v_k are random effects, representing unobserved interviewer and area effects, respectively. The random effects are assumed to follow normal distributions, i.e., $u_j \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$. The variance parameters σ_u^2 and σ_v^2 are respectively the residual between-interviewer and between-area variances in the log-odds of accurate measurement versus measurement with error.

Model (1) was fitted separately for each of the five interviewer observation variables using Markov Chain Monte Carlo (MCMC) estimation in MlwiN (Browne 2009; Rasbash et al. 2009). Model specifications are noted in appendix 2D. For models in which the area effects are not significant, after controlling for household and interviewer effects, model (1) reduces to a simpler two-level model accounting for households within interviewers. For these simpler models, the random effects feature in Stata (command "xtlogit") is used, based on Gauss–Hermite quadrature (Rabe-Hesketh and Skrondal 2012), and the random interviewer variance statistic is designated as rho.

The modeling strategy first explores the random structure, starting with (empty) cross-classified models incorporating both the interviewer and area effects simultaneously but no covariates. Then, groups of explanatory variables are added: the "true" value of the observation (i.e., the Census self-report), the result code, a dummy for survey, then household, interviewer and, finally, area characteristics. Interviewer and area variances are monitored throughout the modeling procedure to understand what type of covariates may explain part of the interviewer or area effect, or both. Observations without significant area effects are interpreted using a two-level model.

The rationale for the selection of covariates is as follows. Accuracy is expected to vary across the different categories of the true value. For example, if there are no children present, the interviewer is not likely to see any children and thus likely to conclude their absence. Whereas if there are children in the household, they could just not be visible at the time of contact and the interviewer would record a false negative observation. The interviewer-assigned result code for the case is entered to test the hypothesis that measurement error is larger for noncontacts or refusals as opposed to cooperative households. The dummy for the six surveys controls for design differences. Household-level information is taken from the Census (e.g., owned/rented, number of adults, indicators of deprivation[5]), and includes basic area information (e.g., urban/rural indicator and region of the country); urban areas and units with restricted access (e.g., flats) are expected to be more difficult to observe correctly. Interviewer characteristics include sociodemographic and work-related information (e.g., education, years of experience, pay grade, other employment) as well as information about attitudes, behaviors, and doorstep approaches (see table E1 in appendix 2E). More experienced interviewers are expected to be more adept at observing respondents' characteristics. In addition, interviewers' characteristics that reflect a willingness to investigate a property more closely (e.g., low score on respecting privacy), good conversational skills (e.g., high scores on keeping a conversation going during contact), and flexibility in their approach and wording are expected to be associated with improved accuracy. Aggregate area-level Census information (e.g., unemployment rate, ethnicity distribution) was included only when the area effect persisted after all other covariates were entered.

2.4 Results

2.4.1 Descriptive statistics

Comparing Census data to the interviewer observations for each case reveals that agreement varies from 87 to 98 percent, depending on the observation (see table B1 in appendix 2B). Type of HU and White were the most accurate (97 percent and 98 percent, respectively), and interviewers were fairly accurate in evaluating the Working and Children observations (both 93 percent). Interviewers had the most difficulty judging accurately the Council observation (87 percent).

[5] The Indices of Deprivation provide a relative measure of deprivation for specific domains, considering poverty, and lack of resources and opportunities, by small area. Areas with higher scores have a higher proportion of deprived people. The domains are calculated from indicators specific to that aspect of deprivation. For example, the Education domain takes into account the average scores of students on various standardized exams, absentee rates, proportion of students not entering higher education, and proportion of adults with no or low qualifications (Department for Communities and Local Government 2011).

A closer look at the direction of the error (i.e., false positive versus false negative percentages) seems to indicate that interviewers more often chose the population mode when they were unsure. For example, when interviewers made an incorrect Working observation, they more often recorded that there was a working adult in the household (false positive rate of 4.0 percent compared to a false negative rate of 3.2 percent). The exception to this pattern is the observation of council properties, where interviewers tended to overestimate the number of council-owned properties.

As expected, interviewers were better at correctly observing the household characteristics for cooperative cases, compared to refusals and noncontacts (see figure 2.1). The accuracy of the Working and Children observations, which were made only for contacted cases, was lower if the household refused. For the Type of HU observation, the lower accuracy for noncontacts is pronounced (92.7 percent compared to 97.0 percent overall).

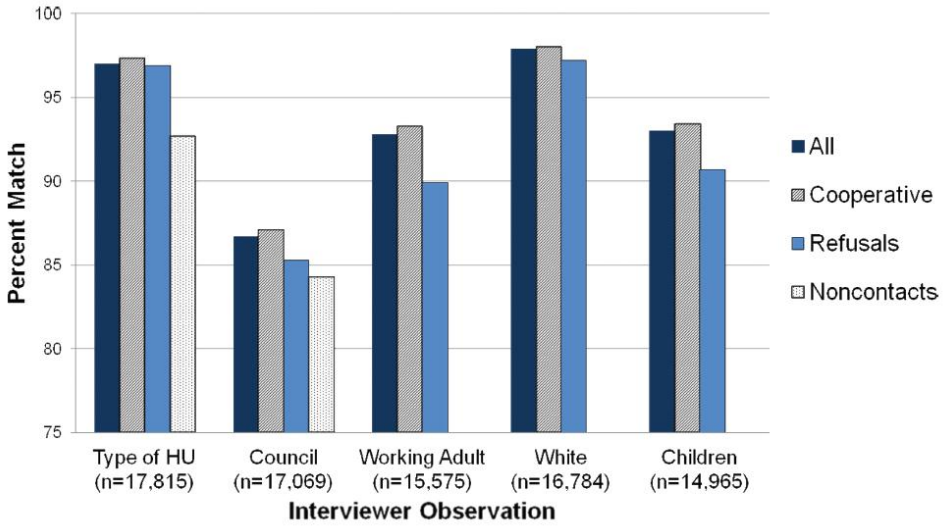


Figure 2.1. Percent Match for Each Interviewer Observation When Compared to Data Obtained from the Census, Overall and by Result Code

2.4.2 Multilevel models: Overall results

To examine the effect of interviewers and areas on the accuracy of each observation, groups of covariates were added to cross-classified multilevel models in a stepwise fashion, from the empty to the final model. The interviewer effects are significant in the empty models for all observations. Except for Working, the variance due to interviewers remains significant for all observations until the interviewer characteristics are added. When these are added, the covariates fully explain the interviewer effect for the Type of HU and Children observations but the interviewer effect remains significant for the Council and White observations. Considering next the area effects, once the first covariates were introduced into the empty models for all observations, the area effect was not significant except for Type of HU and Council. In the model for Type of HU, the area effect is explained by the true value, result code, and household and interviewer characteristics, but it remains significant for the Council observation even after all of these variables as well as area characteristics are included in the model (see appendix 2D).

Given that the effect of area is fully explained for four of the five interviewer observations (Type of HU, Working, White, and Children), simpler two-level models can be used for these four observations (see table 2.2). When two-level models are used for all five observations, there is evidence of significant interviewer influences on the level of measurement error for four of the five observations—Type of HU, Council, White, and Children—as indicated by rho. (Because the two-level and cross-classified models for Council lead to the same conclusions, the cross-classified model is presented only in table F2 in appendix 2F.)

There are a number of similarities across the models. For the three observations that depend on contact (Working, White, and Children) and Council, interviewers are significantly less likely to correctly observe the refused households than the cooperative ones. This is probably because interviewers who receive a refusal have less time with and less access to the household to make the observation.[6] Visibility, as a trait of the household, may also be a factor, with cooperative households displaying observable characteristics more openly than uncooperative households.

The result code is also a significant predictor of accuracy for Type of HU, but here households coded as noncontacts are less likely to be accurately observed than are cooperative households. Because neither Council nor Type of HU requires contact to complete the observation, and both are basically an observation of the outer structure, this is an interesting finding. Assuming interviewers visited the address, contact may be necessary to correctly identify the type of structure (e.g., a building of flats that looks like a single-family house from the outside).

The results show that housing unit structure affects accuracy for almost all observations. For Type of HU, Working, and White, if the structure is a house—as opposed to a flat or other unit—interviewers are more likely to make an accurate observation. This result supports the explanation above, that the Type of HU observation may be problematic when flats can be mistaken for houses. For the Working and White observations, people living in a house are probably more visible than those in flats, making these observations easier. For the Council observation, the result is different; interviewers can more easily identify flats as belonging to a council estate than houses. Surprisingly, housing unit structure has no significant effect on the interviewers' ability to correctly evaluate whether there are children in the household or not, because it was expected that interviewers could more easily observe children in houses than in flats.

Ownership of the property is a significant predictor of accuracy in all models except Children, but the direction varies. For Type of HU, Working, and White, interviewers are more likely to observe these characteristics correctly for owned properties. Because ownership is correlated with the housing unit structure (owners are more likely to own houses; non-owners are more likely to live in flats), the interpretation of these results is similar to that of housing unit structure, discussed above. However, in the case of the Council observation, ownership has a negative relationship with accuracy, showing the direction of error—some owned properties are mistakenly observed as council houses. As detailed earlier, errors in Council may not be an error of the interviewer but rather be due to differences in the interviewer

[6] Note that since the observations were recorded on a paper form, and the details of the interviewers' training are not known, one cannot be exactly certain as to when the observations were made. Observations may have been recorded at first call or contact, or after all contacts.

observation form and the Census question (the address may be on a council estate, as the observation asks, but individual units could also be owned).

Although interviewer experience and age was expected to predict accuracy, these are not always significant. One or both of the characteristics are significant when predicting the accuracy of the Type of HU, Working, and Children observations, but neither is significant for Council or White. It is not obvious why the characteristics are significant for this group of observations. However, when the relationship is significant, the oldest interviewers (60+ years old) are less likely to observe the characteristic correctly compared to other age groups, and the most experienced interviewers (9+ years) are more likely to be accurate compared to at least one of the less experienced groups. Despite the large number of interviewer attitudes available from the interviewer survey, very few are significant predictors of agreement and there is no consistency across models. If more of these variables were significant and the findings consistent, the results could have informed improvements to interviewer selection and training.

2.4.3 Multilevel models: Specific model results

In addition to the predictors of accuracy common across all models, I highlight a few predictors unique to the individual models. In the model predicting accuracy of the Type of HU observation, the significance of the London indicator illustrates the difficulty in making this observation in urban areas. The negative coefficient of the true value in the model predicting the accuracy of the Council observation shows that council houses are likely to be missed or underestimated, contrary to the slight overestimation reported in the descriptive results (not controlling for other factors). The model predicting the accuracy of the Working observation, the only model with neither significant interviewer nor area random effects, finds that interviewers are more likely to correctly observe a working adult than a non-working adult. In addition, household composition covariates (number of children and adults) are significant predictors of accuracy. As the White observation has the highest level of accuracy and a high prevalence in the population (see table B1 in appendix 2B), there is very little variation to explain in the model. Consequently, few predictors of accuracy are significant, but interviewers do tend to err on the side of recording a household as white.

Although the model of the accuracy of the Children observation finds that interviewers underreport the presence of children, accuracy of this observation is improved if young children (0–4 years old) are present in the household. This is probably because younger children are more likely to be home or have “child paraphernalia” visible than older children, making them easier to observe correctly. The significance of the number of adults, with more adults reducing the likelihood of correct observation, may be because it is difficult to classify children who are on the boundary of being an adult or a child.

Table 2.2. Coefficients and Significance for the Final Two-Level Models, with Random Interviewer Effects, Predicting the Accuracy of Each Observation

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	p value	Coefficient	p value	Coefficient	p value	Coefficient	p value	Coefficient	p value
Result Code										
Cooperation	--	--	--	--	--	--	--	--	--	--
Refusal	-0.18	0.132	-0.12	0.048	-0.56	0.000	-0.43	0.002	-0.64	0.000
Noncontact	-0.62	0.000	-0.12	0.301	--	--	--	--	--	--
Area										
London[1]	-0.50	0.000	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Household Char.										
House	--	--	--	--	--	--	--	--	n.s.	n.s.
Flat	-1.89	0.000	0.14	0.047	-0.17	0.064	-0.44	0.013	n.s.	n.s.
Caravan, Other	-2.62	0.000	2.15	0.035	0.44	0.547	--	--	n.s.	n.s.
Own	0.53	0.000	-0.30	0.002	0.76	0.000	0.91	0.000	n.s.	n.s.
Rooms	0.08	0.040	0.23	0.000	-0.08	0.001	n.s.	n.s.	n.s.	n.s.
Council House	0.29	0.038	-0.71	0.000	0.23	0.035	0.61	0.001	n.s.	n.s.
Lowest Floor 1 or 2	-0.73	0.000	n.s.	n.s.	n.s.	n.s.	-0.42	0.043	n.s.	n.s.
Working Adult	0.32	0.003	-0.22	0.001	0.86	0.000	n.s.	n.s.	n.s.	n.s.
0 Cars	n.s.	n.s.	--	--	--	--	n.s.	n.s.	n.s.	n.s.
1 Car	n.s.	n.s.	0.08	0.193	-0.25	0.002	n.s.	n.s.	n.s.	n.s.
2 Cars	n.s.	n.s.	0.68	0.000	0.06	0.602	n.s.	n.s.	n.s.	n.s.
3+ Cars	n.s.	n.s.	0.63	0.000	0.19	0.339	n.s.	n.s.	n.s.	n.s.
1 Adult	--	--	--	--	--	--	n.s.	n.s.	--	--
2 Adults	0.15	0.164	-0.13	0.027	-0.20	0.018	n.s.	n.s.	0.04	0.679

Table 2.2. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
3 Adults	-0.06	0.764	-0.49	0.000	-0.62	0.000	n.s.	n.s.	-0.70	0.000
4+ Adults	-0.59	0.014	-0.85	0.000	-0.14	0.453	n.s.	n.s.	-1.22	0.000
No Children	n.s.	n.s.	--	--	--	--	--	--	--	--
1 Child	n.s.	n.s.	-0.08	0.300	0.00	0.971	-0.28	0.075	-2.91	0.000
2 Children	n.s.	n.s.	-0.21	0.009	0.24	0.030	0.15	0.405	-1.38	0.000
3+ Children	n.s.	n.s.	-0.50	0.000	0.12	0.400	0.77	0.003	-0.70	0.000
Child 0-4 years	--	--	--	--	--	--	--	--	1.11	0.000
All White	n.s.	n.s.	n.s.	n.s.	--	--	3.36	0.000	--	--
Mixed Race Only	n.s.	n.s.	n.s.	n.s.	-0.71	0.224	--	--	-1.19	0.085
Asian Only	n.s.	n.s.	n.s.	n.s.	-0.31	0.158	--	--	0.20	0.444
Black Only	n.s.	n.s.	n.s.	n.s.	-0.66	0.002	--	--	-0.71	0.004
Chinese/Other Only	n.s.	n.s.	n.s.	n.s.	-1.06	0.001	--	--	-1.38	0.000
Mixed HH	n.s.	n.s.	n.s.	n.s.	-0.61	0.000	--	--	0.21	0.297
Total Calls to HH	n.s.	n.s.	n.s.	n.s.	-0.03	0.044	n.s.	n.s.	n.s.	n.s.
Interviewer Char.										
<39 years old	0.41	0.077	n.s.	n.s.	0.32	0.049	n.s.	n.s.	n.s.	n.s.
40- 49 years old	0.47	0.009	n.s.	n.s.	-0.04	0.699	n.s.	n.s.	n.s.	n.s.
50-59 years old	0.28	0.032	n.s.	n.s.	0.14	0.093	n.s.	n.s.	n.s.	n.s.
60+ years old	--	--	n.s.	n.s.	--	--	n.s.	n.s.	n.s.	n.s.
<1 year experience	-0.26	0.151	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.33	0.013
1-2 years experience	-0.30	0.046	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.03	0.789
3-8 years experience	0.02	0.887	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.16	0.120
9+ years experience	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--

Table 2.2. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
Interviewer Survey										
Confident	n.s.	n.s.	n.s.	n.s.	-0.15	0.025	n.s.	n.s.	n.s.	n.s.
Keep Conversation Going-Str Disagree	n.s.	n.s.	0.54	0.023	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Refusal Affects Behavior:										
Always/Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.15	0.690	n.s.	n.s.
Sometimes/Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.35	0.011	n.s.	n.s.
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--	n.s.	n.s.
Refusal Affects Behavior:										
Always	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.36	0.565
Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.65	0.009
Sometimes	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.09	0.517
Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.04	0.704
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Refusal Affects How I Feel:										
Always	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.32	0.325
Frequently	1.04	0.001	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.03	0.857
Sometimes	0.89	0.002	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.15	0.262
Rarely	0.72	0.012	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.29	0.032
Never	0.88	0.004	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Guess Type of People in Home:										
Always	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.39	0.000
Sometimes	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.24	0.033

Table 2.2. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.14	0.282
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.54	0.033
Ask to Enter Home:										
Always	0.58	0.030	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Frequently/Sometimes/ Rarely	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Never	0.13	0.239	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Use Wide Variety of Approaches-Str Agree	-0.37	0.008	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Final Rho (se)	0.033 (0.020)		0.076 (0.011)		0.004 (0.009)		0.062 (0.026)		0.024 (0.012)	
95% CI	[0.010, 0.103]		[0.058, 0.100]		[0.000, 0.213]		[0.027, 0.138]		[0.009, 0.062]	

n.s. = not significant for the accuracy that observation and therefore not shown in the final model

[1]Other area variables controlling for the region of the country are not shown. Other control variables not shown are the deprivation indicators and indicators for the six surveys. See table F1, appendix 2F, for these results.

2.5 Discussion

This analysis examined the level of measurement error in five interviewer observations that were asked about on the 2001 UK Census. Correlates of accuracy were examined with the aim of determining common covariates that affect accuracy. To broadly summarize the findings, the agreement between the observations and the Census reports is generally high. This implies that the interviewer observations analyzed suffer from minimal measurement error, resemble true values, and are, at least in principle, usable for further analysis. Measurement error in the observations is affected by visibility (e.g., as indicated by the type of housing unit structure) and the level of interviewer-respondent interaction (e.g., as indicated by result code). Although the validity is satisfactory overall, there are some signs of variable reliability across the observations. This is indicated by the inconsistent influence of individual interviewer characteristics (experience and age) and the remaining unexplained interviewer variance in some models (though the two-level models suggest that some of this variance can have an area component). Interviewers' answers to questions about attitudes and behaviors do little to explain interviewer variance.

This analysis is limited because of the Census form questions and data. Although there are potential discrepancies between the interviewer observation and Census questions, the analysis included variables in the models to reconcile this (e.g., young children predicting the accuracy of observing any child less than 16 years old). Also, any potential measurement error in the Census data is a minor concern for the variables analyzed. The results provide evidence of the influence of household and interviewer characteristics on observational accuracy, which may inform field practices to improve accuracy. The effects of characteristics like housing unit structure and ownership on accuracy indicate that more effort may be necessary when observing rental properties and flats. However, other findings, such as the effect of interviewer experience, require more study to understand the mechanism(s). For example, experienced interviewers may be more accurate because they are more familiar with the areas they work with or are more comfortable soliciting proxy information from neighbors. Therefore, additional data might help disentangle these possibilities. Any resulting recommendation would depend on the mechanism.

One way to improve accuracy in general, and also possibly remove the effect of more experienced interviewers, is more rigorous interviewer training. Some surveys, such as the Los Angeles Family and Neighborhood Study (L.A.FANS), prioritize the accuracy of interviewer observations and develop special training (see Casas-Cordero 2010, pp. 74–75). The NSFG successfully experimented with providing interviewers with visible household and person characteristics that predicted the characteristic being observed to improve accuracy (West 2010). These solutions, however, involve more effort on the part of the interviewers, researchers, and managers, and may come at additional cost.

There was evidence of differential measurement error by result code. It makes sense that interviewers have more information on cooperative and contacted cases. However, with these data it is not clear if this result is confounded with the point in time when the observations were made, since a paper-based form allows for little quality control. This quandary highlights the value of installing firm protocols for the collection of interviewer observations and checking that interviewers follow them. Differences between the quality of the observations for cooperative and non-cooperative households (or any other systematic

measurement error in the observations) are likely to have an adverse effect on statistical adjustments (West 2013b) and decisions based on the observations.

Additional lessons from the analysis underscore practical cautions for designers of interviewer observations. First, it may not be advantageous to collect observations that interviewers are not able to easily observe. The Children observation suffered from high missing-data rates, indicating difficulty in observing children. In addition, the accuracy analysis reveals that young children are easier to observe than older ones. If an observation of young children is sufficient, then collecting this more precise question is likely to yield higher-quality observational data. Second, as with questionnaire items, observation questions should accurately capture the construct of interest and be understood consistently by interviewers. Although the Council observation asked if the property was on a council estate, researchers may more specifically need to know if the property is in fact owned by the local authority. Finally, when deciding what observations to collect, the intended application should be considered. Observations meant to correct for nonresponse bias should be highly correlated with both survey participation and the survey outcomes of interest in order to be effective (Little and Vartivarian 2005). Using these criteria, the most effective observations will vary depending on what survey outcome is being adjusted.

This analysis is the first step to understanding and reducing the measurement error in interviewer observations. The observations explored are similar or identical to some of those collected by the large-scale studies mentioned in the introduction (e.g., type of housing unit in the ESS). The findings of only small levels of measurement error are good news for these and other surveys using similar interviewer observations. However, for other observations, such as whether the respondent is sexually active or not, as collected in the NSFG, the level of measurement error and the mechanisms behind it may or may not be similar to what is found here. Therefore, the findings cannot be safely extrapolated to all possible observations.

Chapter 3: Comparing Interviewer Observations to Commercial Data

Note: This is a pre-copy-editing, author-produced PDF of an article accepted for publication in *Public Opinion Quarterly* following peer review. The definitive publisher-authenticated version

Sinibaldi, Jennifer, Mark Trappmann, and Frauke Kreuter. *2014*. "Which is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations?" *Public Opinion Quarterly*, 78:2.
doi: 10.1093/poq/nfu003

is available online at: http://poq.oxfordjournals.org.

3.1 Introduction

Users of survey data need high-quality auxiliary data on both respondents and nonrespondents in order to effectively correct for nonresponse bias. The most useful auxiliary data variables are highly correlated with the outcomes of interest and the propensity to respond (Little and Vartivarian 2005). In the absence of rich frame data, researchers have two primary options for auxiliary data that may fit these criteria. First, commercially available data on small areas or households can be purchased and linked to the survey data. Second, paradata in the form of specially designed interviewer observations specific to an area, household, or person can be recorded during the data collection. Costs and errors associated with both options must be weighed.

Commercial data provide regional and household information on characteristics such as household composition, property types, leisure activities, and purchasing power[7] and can include source data that are either inaccurate or incorrectly processed (Kapteyn and Ypma 2007). Also, if household characteristics are not updated quickly enough when either the characteristics of the occupants or the occupants themselves change, quality will diminish (for an example of this error in establishment data, see Groen 2012). Furthermore, for confidentiality reasons some commercial data vendors, like microm in Germany, do not provide individual-level measurements but instead aggregate data by small clusters of households. This aggregation can introduce uncertainty when using commercial data to predict individual level variables (Biemer and Peytchev 2012). Missing data can also be a problem if some of the sample units cannot be linked to commercial data, possibly introducing selectivity bias (Huynh et al. 2002). Lastly, even if accuracy is high, commercial data may not measure exactly the same concept as the one measured in the survey (Davies and Fisher 2009).

Existing work on the quality of commercial data indicates that these data are inadequate for some purposes. While investigating the coverage of ethnicity designations on a commercial list for the purpose of enriching the frame for sample selection, researchers on the Racial and Ethnic Approaches to Community Health (REACH) project found that the quality was not consistent across ethnic groups and often suffered in more urban areas, especially when the ethnic concentration was diverse and/or impoverished (English et al. 2012). When

[7] Two examples of commercial data vendors are Experian and microm. See www.experian.com/marketing-services/consumer-data.html and www.microm-online.de for information about the range of commercial data provided by these companies.

comparing the self-reports of Knowledge Networks panelists to auxiliary data, DiSogra et al. (2010) also found low accuracy in the auxiliary data for ethnicity, as well as low correlations (ranging from 0.261 to 0.634) between the two data sources on eight household characteristics. The Survey Research Center at the University of Michigan reviewed the quality of its commercial data to determine if the cost of renewing the contract required to access the data was justified. The report notes problems with duplicate records for the same address (10%), items missing data (e.g., the presence of children indicator and household size were both missing for 9% of the records), and dissimilarities between the distributions of some of the commercial variables (e.g., age and household size) when compared to Census data. Although the shortcomings were numerous, when data were available, the indicators seemed to improve the strength of models predicting household eligibility (Hubbard and Lepkowski 2009).

Commercial data have also been examined by government statistical agencies. Employing databases generated from the customers of fourteen companies, the UK's Office for National Statistics explored the usefulness of commercial databases for producing specific population statistics. Some advantages were cited, especially regarding the level of detail in the purchase data, but because of coverage error, regional biases, and various other inaccuracies, these data could not be used to calculate precise population statistics (Dugmore 2010). Similarly, the US Census Bureau's investigation of the quality of both government administrative data and commercial data reports that these auxiliary sources are currently not of high enough quality or coverage "to replace a traditional census" (Rastogi and O'Hara 2012, xii). However, auxiliary data could be useful for the enhancement of Census data and cost reduction in follow-up efforts (Rastogi and O'Hara 2012).

An alternative to commercial data would be interviewer observations of area, household, and person characteristics recorded during the survey data collection, the quality of which has also recently come under investigation. Since the accuracy of the recorded characteristics is dependent on the observational abilities of the interviewers, how readily observable the characteristics are, and the interviewers' correct interpretation of what they see, these data can suffer from interviewer, area, and household effects (see West and Sinibaldi 2013 for an in-depth description of factors that may affect the quality of interviewer observations). Work by West (2013a) on the National Survey of Family Growth (NSFG) found observations of children and sexual activity to be 72% and 78% accurate respectively, compared to self-reports in the survey data. Sinibaldi et al. (2013) found that broadly categorized household observations, such as housing type and employment status, were very accurate (between 87 and 98%), when recorded. However, the accuracy did vary by household characteristics and the level of interaction between interviewer and respondent. Additionally, some observations (e.g., the presence of children) suffered from notable levels of missing data that, if not missing at random, would result in biased estimates when used for adjustment (West and Little 2013). Finally, mismatch between the construct of interest and the instructions in the observation question impairs the ability of observations to provide useful data (e.g., the public housing observation in Sinibaldi et al. 2013). To summarize, many of the potential weaknesses of interviewer observations (e.g., missing data, variations in accuracy by household characteristic, etc.) are similar to those outlined for commercial data.

Across all of these studies, the quality of the auxiliary data, be it commercial data or interviewer observations, is likely to vary by country and data collection agency.

Nonetheless, reports by each organization within each country consistently indicate that the quality of these data falls short for the intended purpose. This analysis acknowledges the inaccuracies of auxiliary data and, given the shortcomings, evaluates which is *more* accurate in the context of a German economic survey. Although no published research is available on the quality of commercial data or interviewer observations in Germany, the data are used for nonresponse analyses and weighting (Schräpler et al. 2010; Trappmann 2011), making the analysis relevant.

The analysis that follows uses respondent-reported survey data to determine which data source—interviewer observations or commercial data—shares more information with the survey data and therefore, is of higher quality and a better investment of the survey budget. Previous research has examined interviewer observations and commercial data separately, making the current study the first to compare both sources within the same analysis. Furthermore, since it is possible that a single source alone may not be of sufficient quality but may improve the results when combined with other auxiliary data (see the work of the federal statistical agencies noted above), the analysis will assess the performance of the interviewer observations and commercial data when used together. Having evaluated one of the two criteria for a good adjustment variable, the auxiliary data (or combination of data) determined to be most predictive of the survey outcomes will be the best choice for nonresponse adjustment, on this dimension.

Balancing the conclusions about quality is the element of cost for each data source. Researchers are looking to these data to treat a problem created by falling response rates, and purchasing auxiliary data is one of several options (other examples are extension of the field period or special interviewer training on refusal conversion) that could be funded by the survey budget to address this problem. This analysis assumes that data must be purchased for nonresponse adjustment and aims to show researchers charged with deciding how to spend their budget for nonresponse adjustment which source is the better investment. Although specific costs for the collection of interviewer observations and the purchase of commercial data are rarely shared by agencies, the costs of these data are disclosed and the impact of cost in light of the conclusions from the data analysis is discussed.

3.2 Data

This analysis incorporates three data sources: survey data that provide the dependent variables of unemployment benefits (UB) and income; interviewer observations as one source of auxiliary data; and commercial data as a second source. These three datasets are explained in more detail below.

3.2.1 Survey data

The Panel Study of Labor Market and Social Security (PASS) is an annual survey of German households used to track changes in unemployment and related economic measures in the country. Designed as a dual-frame survey, the study follows households from both the general population (stratified by social status) and those known to receive, or to have received, unemployment benefits[8] (see Trappmann et al. 2010 and Bethmann and Gebhardt

[8] The specific unemployment benefit studied is called “UB II” or “Arbeitslosengeld II.” It is a means-tested benefit for households with insufficient income. At least one person in the household must be between 15 and 64 years old and able to work a minimum of 15 hours a week. People who are “under-employed” or “working poor” qualify if they meet the requirements. Other recipients may be active in labor market policy programs, seeking education, or out of the labor force (e.g., single parents of small children). UB II must be distinguished from UB I, which is an insurance payment paid in the first year of unemployment to people in need who are not able to work or older

2011 for details about PASS). In wave five, collected from February to September 2011, the panel sample was refreshed with additional households from new primary sampling units for both the general population (henceforth, this general population refreshment sample is called "GP") and the unemployment benefit recipients (henceforth, this refreshment sample of benefit recipients from new *regions* is called "UBR"). A third refreshment sample was drawn from the households in the original sampling points which began receiving benefits in the year between the last and current sample selection (July 2009 and July 2010), as is standard procedure for each wave (henceforth called "UBN" for those households *new* to UB in the last year). The data used for this analysis comprise these three refreshment samples only, which include 6,237 households from the general population and 8,220 from the unemployment registry (5,428 UBR; 2,792 UBN). The response rates were 25.2%, 25.7%, and 28.2% respectively (using AAPOR RR1, AAPOR 2009). These rates are slightly lower than those typically achieved in German face-to-face studies (Schnell 2012) but consistent with PASS response rates for other waves (see Kreuter et al. 2010a for wave one; see West et al. 2013 for waves 2 and 3).

The survey questions analyzed refer to the receipt of unemployment benefits and household income. The self-report of whether anyone in the household is currently receiving unemployment benefits is derived from a series of spell-duration questions which were categorized as: on UB, not on UB, or missing. Since a valid response to these items is necessary for the analysis, the 0.2% of cases missing data on UB were excluded (see table 3.2). The total net income of the household was calculated from a series of detailed income questions, providing prompts for all possible sources of income. Cases with missing income data (2.1%) were excluded from the analysis (see table 3.2). The remaining income responses were adjusted for household size, using the OECD transformation (see Gebhardt et al. 2009, 87).

To allow for comparison with the interviewer observations, the continuous values of self-reported income were divided into three categories: low, medium and high. To do this, the distribution of OECD-adjusted income for all respondents in 2011, weighted for selection and nonresponse (see Trappmann 2011), was divided into thirds. This process provided the cut points of low, medium, and high income in the population, which could then be used to classify the households in the three refreshment samples analyzed as high, medium, or low income[9]. Due to oversampling of low income households in the general population and the two samples specifically targeting households on UB, the distribution of the income variable in the data is disproportionately low compared to the population (West et al. 2012).

3.2.2 Interviewer observations

Interviewers collected observations specially designed to closely resemble the UB and income survey questions for all contactable CAPI cases. Every household in the three refreshment samples was assigned to CAPI data collection initially. If a household refused or could not be contacted, the case was assigned to CATI. Therefore, any CATI-only cases and cases that were never sent to the field do not have observations and are not included in the analysis (see table 3.2). In addition, 4.8% of all cases sampled for wave five (not just refreshment) with observations had two sets of observations recorded due to reassignment

than 65 years old. Households can receive both benefits if their UB I claim is low enough to still qualify them for UB II.

[9] Low income households made less than 1067 Euros per month. Mid-income households made 1067 or more but less than 1667 Euros per month, after adjustment.

of the case. Duplicates were cleaned first by removing observations keyed in from paper forms in favor of the CAPI-entered observations, and then by keeping the observation with the earlier date. All observations were associated with a single interviewer.

The wordings of the interviewer observation questions (translated from German) are:

1. Do you think that the household income of the household living at this address is low, medium, or high, compared to the total population?
2. Do you think that someone in this household is currently receiving unemployment benefits[10]? Yes, No

Interviewers had not collected these observations previously and received brief training, including a training memorandum, on them. Interviewers were instructed to record the observations on a paper address form at the first visit to the address and told not to edit the observations at a later point, even when entering them into the CAPI system. The memorandum emphasized that the observation must be recorded before the interview and that there are no right or wrong answers. It is important to note that an experiment designed to help interviewers make better observations was conducted in wave 5. Interviewers received predictions (made from frame and microm data) of the income bracket and UB status for a random half of their cases (West et al. 2012). The experiment found that providing these predicted outcomes did not improve the accuracy of the interviewer observations, and qualitative interviews revealed that not many interviewers used the predictions when making their observations. Nonetheless, the experimental group is controlled for in the analysis.

Finally, the observations suffer from missing data (2.3% missing in the analysis dataset). These cases are retained by putting them into their own category. However, when the PASS samples are analyzed separately, cross tabs between the observations and the dependent variables indicate that the observations are missing for some but not all of the categories of self-reported income, resulting in “empty cells” that cannot be collapsed with another category. Empty cells are present in the UBR (no cases have missing observations *and* medium or high income) and UBN (no cases have missing observations *and* medium income) samples. Therefore, when modeling self-reported income, cases in the missing category of the observations were dropped for the two UB refreshment samples (see table 3.7).

3.2.3 Microm data

Commercial microgeographic data provided by microm Micromarketing-Systeme und Consult GmbH[11] (henceforth, Microm) can be linked at the address level for approximately 40.8 million households in Germany. Microm data are compiled from multiple sources including government records and surveys and information from postal and telecommunication carriers, and are intended to aid users in defining particular commercial markets (Oemmelen 2012). In this analysis, desireable indicators are those that are theoretically correlated with the self-reported income or UB recipiency at the household level, and therefore only the

[10] The interviewers were specifically asked to observe if someone was on UB II. The interviewers judge receipt of this benefit based on their general knowledge (it is much discussed in Germany) and their experience with PASS households in previous waves. Since UB II is mainly for those in chronic poverty, the observation is less about employment status and more about assessing whether the household is poor.

[11] As noted in the introduction, see http://www.microm-online.de for information about the microm company and the data collected.

indicators compiled at the smallest area level, which is a minimum of five households (average aggregation of 7.5 households) (Oemmelen 2012), are considered. There are 14 indicators at this level, several of which capture similar information. The analysis uses six of these indicators, considering known correlates of unemployment benefit recipiency (such as age, migration and the presence of children; see Achatz and Trappmann 2011, Fuchs 2012 and Riphahn et al. 2013) and income in Germany (such as education and unemployment; see Biewen 2006 and Bundesministerium für Arbeit und Soziales 2013). Also considered were correlations between the indicators and the dependent variables seen in the data and multicollinearity between the indicators. The indicator most closely aligned with the survey outcomes is the social status of the household, which is based on education and income (Schräpler et al. 2010). Other indicators are housing type, family composition, proportion of household members under 30, proportion of foreigners (derived by examining first and last names (microm 2013)), and the migration into and out of the micro-area of 5+ households (called mobility).

Table 3.1. Description of Microm Variables Used in the Analysis with Labels for the Categories

Variable name	Description
House type	Concentration of family homes
	(1) 1-2 family home on streets with a homogeneous building structure
	(2) 1-2 family home on streets with mixed building structure
	(3) 3-5 family home
	(4) 6-9 family home
	(5) Apartment block with 10-19 households
	(6) High rise buildings with 20+ households
	(7) Households combined with mostly commercial space
Mobility	Measure of households moving in and out
	(1)Very strongly negative rate - moving out
	[…]
	(9) Very strongly positive - moving in
Under 30	Percent of heads of household under 30 years old
	(0) Up to 5%
	[…]
	(9) Over 50%
Foreign	Proportion of foreigners
	(1) No foreigners
	[…]
	(9) Highest proportion
Family type	Composition of families
	(1) Mostly single person households
	[…]
	(9) Almost exclusively families with children
Status	Status (wealth & prominence) of household
	(1) Lowest social status
	[…]
	(9) Highest social status

Table 3.1 presents a description of the six indicators with labels for the endpoints of the scale. In the analysis dataset, 4.5% of households have missing Microm data (4.7% in GP, 4.8% in UBR, and 3.6% in UBN) because the addresses could not be matched. Missing data is kept in the analysis as a category because it is an important aspect of using auxiliary data that should be considered. All categories of the indicators are retained, collapsing only when necessary. As with the interviewer observations, the problem of "empty cells" that cannot be collapsed with another category results in reductions in case base for particular samples. See appendix 3A for additional details on missing Microm data.

3.2.4 Analysis dataset

As noted earlier, the analysis is limited to respondents with valid responses to the survey questions of interest. In addition, two households that were located in an entirely commercial zone were removed from the analysis, and one case was lost due to missing data on the East Germany indicator. This results in a final case base of 3,213 households for the analysis (see table 3.2). Within each of the samples, 22-24% of the cases selected for the survey are included in the analysis, resulting in 1,377 cases for GP, 1,176 for UBR, and 660 for UBN. This is 85%-89% of the respondents for these samples. As mentioned above, these case bases are further reduced for some models because of empty cells in the auxiliary data that appear when analyzing a specific survey outcome (explained in appendix 3A for Microm and evident in tables 3.6 and 3.7).

Table 3.2. Size of the Analysis Sample after Each Step of Data Cleaning, Shown for the Full Dataset and Each Sample Separately

		Remaining cases			
	Cases lost	**Overall** **n=14,457**	**General Population Refreshment (GP)** **n=6237**	**UB Refreshment, new regions (UBR)** **n=5428**	**UB Refreshment, new in the last year (UBN)** **n=2792**
Reduced sample size after removing:					
Cases not contacted via CAPI	2322	12135	5546	4346	2243
Unit nonrespondents (no survey data)	8846	3289	1431	1188	670
Cases missing values on both income and UB self-reports	3	3286	1429	1187	670
Cases missing values on UB self-report only	5	3281	1429	1182	670
Cases missing values on income self-report only	65	3216	1379	1176	661
House in a business district, according to Microm data	2	3214	1378	1176	660
Address cannot be classified as East or West Germany	1	3213	1377	1176	660
Final case count for analysis		3213	1377	1176	660
Percent of respondents in analysis		86.8	89.4	85.0	84.8

3.3 Methods

To answer the research question of which type of auxiliary data is more predictive of the outcomes of interest, the two survey outcomes (UB and income) are analyzed separately. In the multivariate analysis, although interpretation of the coefficients is not necessary to accurately answer the research questions, model development is important. Appendix 3B details the modeling decisions made when assessing the hierarchical nature of the data (i.e., households nested within interviewers and areas) using cross classified and multilevel models, as well as exploring interaction effects. These investigations concluded that random interviewer and area effects are not significant and simple logistic (for UB) and ordered logit (for income) models are appropriate, when run separately for each sample. Within a sample, the significance of the parameters associated with the predictors from the auxiliary data, in a model with both auxiliary data sources included as predictors, should indicate which data source is the more powerful predictor of the outcome. To verify this assessment, cross validation is conducted.

In the cross validation, the cases are divided into five equally sized random subsamples (using the same seed) within each of the three PASS refreshment samples. Models using only interviewer observations, only Microm indicators, or both, are fit to data from four of the five subsamples. The coefficients from these models are applied to the data in the fifth subsample (i.e., the validation subsample) to calculate the predicted probabilities for each case in this subsample only. For UB, one probability predicting "on UB" is generated. For income, a probability is generated for each category. This process of using four-fifths of the sample to calculate predicted probabilities for the remaining fifth is repeated four more times, using a different subsample as the validation subsample each time (see figure 3.1).

An assessment of the accuracy of the model is calculated by taking the squared difference between the predicted probability for each case in the validation subsample and the survey value. For UB, there is only one predicted probability per case and the survey value is either 0 or 1. For income, the difference is calculated using the predicted probability for the category reported in the survey data and 1. The mean of these squared differences (called scores in figure 3.1) is calculated for each validation subsample. Paired t-tests (two-tailed, α=0.05) using all cases in a validation subsample are then used to compare the scores of the models using a single data source (interviewer observations or Microm) to each other as well as to compare each single source model to a model using both sources. Since this comparison is done for each validation subsample, five sets of comparisons for each survey outcome are created for each PASS sample. The cross validation technique illustrates the results observed during the modeling process, allowing us to more confidently conclude which type of auxiliary data is more predictive of each survey outcome in these data, and therefore, a better investment of survey budget.

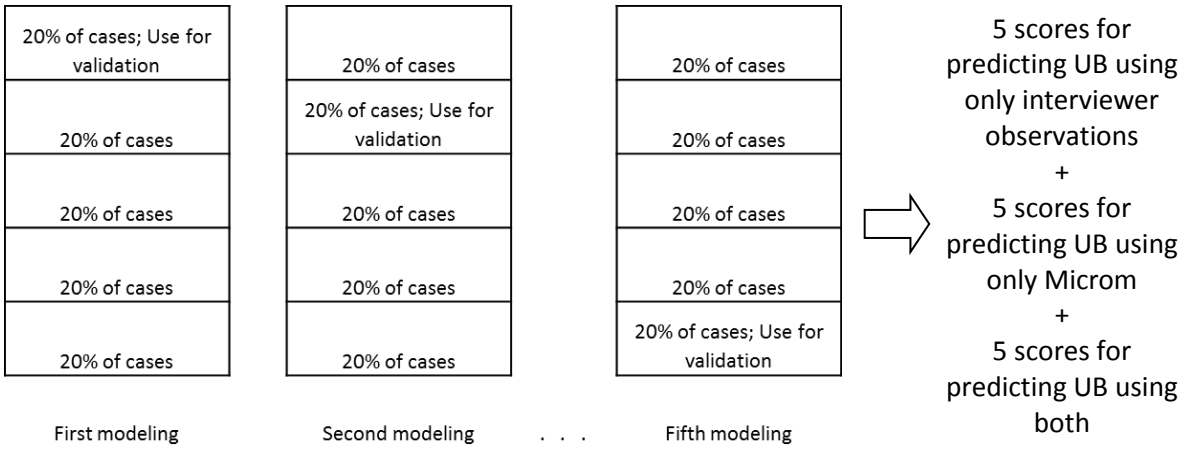


Figure 3.1. Diagram Describing Cross Validation for a Single PASS sample (e.g., General Population Refreshment) for One of the Survey Outcomes (e.g., UB)

3.4 Results

3.4.1 Descriptive analysis

To gain some perspective on the explanatory power of the interviewer observations used in the analysis, the observations were compared to the survey values. Tables 3.3 and 3.4 show the distribution of the unemployment benefit recipiency and income interviewer observations and the percent of each response that corresponds with the survey responses. For both, the observation is significantly correlated with the survey variable (UB $X^2(2) = 797.0$; income $X^2(6) = 763.7$). The overall agreement for UB is 74.3%, which is moderate given the higher levels of agreement found in Sinibaldi et al. 2013 (but similar to the agreement found by West 2013a). Using these same studies for comparison, the 55.8% agreement for income appears to be poor. The distributions of the observations and the percent agreement with UB and income for each PASS sample (GP, UBR and UBN) can be found in appendix 3C (tables C1 and C2).

Table 3.3. Frequency of the Interviewer Observations of Unemployment Benefit Status and the Percent within Each Observational Category that Corresponds with the Self-Reported Value from the Survey

		Unemployment Benefit: Self-reported	
UB: Interviewer Observed		**On UB** n=1866	**Not on UB** n=1347
	(N)	(%)	(%)
On UB	1906	72.8	27.2
Not on UB	1234	21.9	78.1
Missing	73	43.8	56.2

Table 3.4. Frequency of the Interviewer Observations of Income and the Percent within Each Observational Category that Corresponds with the Self-Reported Value from the Survey

		Income: Self-reported		
Income: Interviewer Observed		**Low** n=1961	**Medium** n=684	**High** n=568
	(N)	(%)	(%)	(%)
Low	1511	82.3	13.7	4.0
Medium	1362	45.2	29.2	25.6
High	267	19.1	24.7	56.2
Missing	73	69.9	17.8	12.3

The distribution of the Microm indicators overall and across the three samples can be found in appendix 3D (table D1). Since the Microm indicators do not specifically measure income or unemployment benefit recipiency, evaluations of agreement are not appropriate. Instead, tests of independence and the strengths of association are presented in table 3.5. All Chi-square tests are significant at $p<0.001$ and the strengths of all associations appear to be weak to moderate, varying from 0.16 to 0.30 for UB and 0.13 to 0.26 for income. The weakest relationship is with the percent of householders under 30, and the strongest is with household status.

Table 3.5. Tests of Independence and Measures of the Strength of Association between the Microm Variables and the Self-Reported Values from the Survey

	Unemployment Benefit n=3067			**Income (categorized)** n=3067		
	Chi square	**df**	**Cramer's V**	**Chi square**	**df**	**Cramer's V**
House type	231.1	7	0.268	281.8	14	0.209
Mobility	153.8	9	0.219	178.5	18	0.167
Under 30	84.9	10	0.163	104.2	20	0.127
Foreign	99.9	9	0.176	131.9	18	0.143
Family type	228.3	9	0.267	258.5	18	0.201
Status	293.1	9	0.302	432.5	18	0.259

Note: All Chi-square tests are significant at $p<0.001$

3.4.2 Multivariate results

The final logistic models predicting whether someone in the household is on UB are presented in table 3.6 for each of the three refreshment samples. When predicting self-reported unemployment benefit recipiency, the interviewer observation of UB is highly significant and in the expected direction, with those households observed as on UB being more likely to have reported being on UB in the survey compared to households observed as not being on UB. Overall, the income observation is not significantly predictive of unemployment benefit recipiency in the GP and UBN samples[12], and the directions of the coefficients in both UB refreshment samples are the opposite of what one would expect, due to a low number of cases in the high-income category[13]. The Microm indicators vary in their significance across the samples and are not significant overall (using a multiparameter Wald test) in either UB refreshment sample. The indicators do contribute to the model of the GP sample, as shown by the significance of the coefficients and the slightly higher pseudo R^2 value for the model using only Microm data ($R^2 = 0.30$), compared to using only interviewer observations ($R^2 = 0.23$) (see appendix 3E, table E1 for all pseudo R^2 values). Therefore, Microm data seem to better explain UB recipiency in the general population than in the samples specific to past or current UB recipients. Interestingly, the missing data indicators are predictive of UB recipiency for two of the samples. The coefficient indicates that households with missing interviewer observations are more likely to be on UB in the UBR sample. Conversely, the households missing Microm indicators are less likely to be on UB.

Predicting low, medium, or high household income in the survey was explored using an ordered logit model (see table 3.7). Across all samples, both interviewer observations are significant and the coefficients are in the expected directions. That is, if the interviewer observation recorded that a household is on benefits, the household is less likely to have reported high income in the survey than a household not observed to be on benefits; and if an interviewer recorded a household as having high income, the likelihood of reporting high income in the survey is much higher than if the interviewer recorded the household as having low income. As with UB, very few of the Microm indicators are significant, but the status indicator seems to show promise, being significant overall in the GP and UBN samples. Although the pseudo R^2 values for the models using only Microm data are higher than those for the models using only interviewer observations for the two UB refreshment samples[14], the overall significance of these models using only Microm data is weaker, and not significant at all for the UBN sample (see appendix 3E, table E1).

[12] Only 6.6% of the cases in the GP analysis reported being on UB II, weakening the ability of the income observation to predict UB II.

[13] The percentage of cases observed as high income were 1.1% and 2.9% in the analyses of the UBR and UBN samples, respectively. This is compared to 16.8% for the analysis of the GP sample.

[14] Note, the pseudo R^2 is less than 0.10 for both models in both samples.

Table 3.6. Odds Ratios from Logistic Regression Predicting Unemployment Benefit Recipiency for Each PASS Refreshment Sample

	General Population Refreshment (GP) n=1295		UB Refreshment, new regions (UBR) n=1176		UB Refreshment, new in the last year (UBN) n=660	
	Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
Treatment group	1.17	0.597	0.91	0.516	0.66*	0.024
East Germany	1.84	0.120	1.28	0.221	0.60*	0.034
Interviewer Observation UB						
Not on UB	(ref)		(ref)		(ref)	
On UB	11.43**	0.000	4.94**	0.000	6.10**	0.000
Interviewer Observation income						
Low	(ref)		(ref)		(ref)	
Medium	0.62	0.174	1.71**	0.008	1.27	0.309
High	0.53	0.443	3.58	0.123	2.80#	0.060
Missing						
Interviewer Observations	2.42	0.265	5.71**	0.002	0.97	0.966
Microm	n.a.		0.95	0.946	0.23#	0.066
House type						
1-2 family home, homogeneous street	(ref)		(ref)		(ref)	
1-2 family home, mixed	0.21#	0.051	0.68	0.311	0.97	0.938
3-5 family home	2.57	0.150	0.61	0.192	0.84	0.647
6-9 family home	3.20#	0.095	0.75	0.458	0.56	0.149
Apartment block with 10-19 households	2.06	0.313	0.86	0.726	0.87	0.750
High rise buildings with 20+ households	1.27	0.768	0.85	0.723	1.27	0.650
Combined with commercial space	n.a.		1.12	0.896	1.30	0.754
Mobility						
Very strongly negative rate	(ref)		(ref)		(ref)	
Strongly negative rate	0.79	0.620	0.80	0.423	1.71	0.113
Negative rate	0.46	0.169	0.81	0.442	1.48	0.282
Slightly negative rate - moving out	0.40	0.123	0.86	0.632	1.69	0.192
Balanced rate	0.33#	0.089	1.10	0.783	1.49	0.327
Slightly positive rate - moving in	0.13*	0.015	0.87	0.705	1.93	0.146
Positive rate - moving in	0.38	0.191	1.66	0.213	1.71	0.245
Strongly positive rate - moving in	0.08**	0.006	1.02	0.954	2.41#	0.067
Very strongly positive rate - moving in	0.33	0.134	2.20#	0.080	1.00	0.995
Under 30						
Up to 5%	(ref)		(ref)		(ref)	
5% - 10%	0.28	0.260	1.38	0.445	1.16	0.749

Table 3.6. *Continued*

	General Population Refreshment (GP) n=1295		UB Refreshment, new regions (UBR) n=1176		UB Refreshment, new in the last year (UBN) n=660	
	Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
10% - 15%	2.57	0.164	0.89	0.759	0.89	0.795
15% - 20%	0.28	0.105	1.10	0.803	1.08	0.870
20% - 25%	1.34	0.690	1.29	0.495	1.21	0.637
25% - 30%	1.18	0.799	2.19*	0.048	1.00	0.992
30% - 35%	1.12	0.867	1.18	0.671	0.78	0.559
35% - 40%	2.34	0.265	1.29	0.519	0.69	0.423
40% - 50%	2.42	0.124	1.14	0.705	1.06	0.893
Over 50%	1.09	0.896	0.93	0.844	1.15	0.731
Foreign						
No foreigners	(ref)		(ref)		(ref)	
Extremely low proportion	0.89	0.867	1.01	0.980	0.41*	0.025
Very low	0.25	0.155	0.84	0.630	0.61	0.253
Well below average	1.27	0.731	2.15#	0.089	0.40*	0.038
Below average	0.87	0.855	0.91	0.818	0.28**	0.003
Slightly below average	2.01	0.319	1.76	0.147	0.44#	0.051
Average	2.25	0.208	1.66	0.163	0.43*	0.034
Above average	0.50	0.357	1.29	0.481	0.44*	0.038
Highest proportion	0.83	0.780	1.28	0.490	0.35**	0.007
Family type						
Mostly single person households	(ref)		(ref)		(ref)	
Well above average proportion of single person households	0.86	0.786	1.44	0.231	0.75	0.434
Above average proportion of single person households	0.64	0.456	0.92	0.785	0.52#	0.075
Slightly higher than average proportion of single person households	1.04	0.947	0.90	0.729	0.64	0.234
Mixed family structure	3.81*	0.030	0.87	0.671	0.75	0.468
Slightly higher than average proportion of families with children	2.73	0.198	0.99	0.987	0.72	0.441
Above average proportion of families with children	3.52	0.102	1.00	0.997	0.99	0.979
Well above average proportion of families with children	4.31#	0.084	1.07	0.877	0.77	0.605
Almost exclusively families with children	0.98	0.982	0.55	0.251	0.20**	0.009

Table 3.6. *Continued*

		General Population Refreshment (GP) n=1295		UB Refreshment, new regions (UBR) n=1176		UB Refreshment, new in the last year (UBN) n=660	
		Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
Status							
	Lowest social status	(ref)		(ref)		(ref)	
	Very low status	0.19**	0.004	0.83	0.448	0.51*	0.025
	Well below average status	0.54	0.187	0.80	0.404	0.95	0.888
	Below average status	0.18**	0.007	0.67	0.169	0.46*	0.039
	Slightly below average status	0.17**	0.007	1.06	0.876	0.97	0.937
	Average status	0.22*	0.013	1.16	0.678	0.50#	0.073
	Slightly above average status	0.09**	0.005	0.62	0.175	0.99	0.973
	Above average status	0.10*	0.012	1.19	0.692	0.45#	0.084
	Highest social status	0.09*	0.033	0.27*	0.016	0.38#	0.071

# p< 0.10, * p< 0.05, ** p< 0.01

Table 3.7. Odds Ratios from Ordered Logit Regression Predicting Low, Medium, and High Income for Each PASS Refreshment Sample

	General Population Refreshment (GP) n=1377		UB Refreshment, new regions (UBR) n=1134		UB Refreshment, new in the last year (UBN) n=614	
	Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
Treatment group	0.96	0.679	1.03	0.874	1.01	0.660
East Germany	1.03	0.853	0.43**	0.006	1.11	0.722
Interviewer Observation UB						
Not on UB	(ref)		(ref)		(ref)	
On UB	0.69*	0.034	0.62#	0.068	0.48**	0.006
Interviewer Observation income						
Low	(ref)		(ref)		(ref)	
Medium	3.03**	0.000	1.44	0.171	2.37**	0.001
High	6.09**	0.000	4.35*	0.035	1.36	0.614
Missing						
Interviewer Observations	1.60	0.205	n.a.		n.a.	
Microm	1.56	0.359	1.87	0.546	n.a.	
House type						
1-2 family home, homogeneous street	(ref)		(ref)		(ref)	
1-2 family home, mixed	0.98	0.915	2.04	0.184	0.74	0.460
3-5 family home	0.81	0.281	2.48#	0.085	0.65	0.281
6-9 family home	0.73	0.182	2.66#	0.071	0.46#	0.076
Apartment block with 10-19 households	0.78	0.322	3.13*	0.049	0.60	0.297
High rise buildings with 20+ households	0.93	0.809	3.94*	0.037	0.41	0.141
Combined with commercial space	1.14	0.806	n.a.		n.a.	
Mobility						
Very strongly negative rate	(ref)		(ref)		(ref)	
Strongly negative rate	0.82	0.492	1.16	0.718	0.73	0.462
Negative rate	1.12	0.708	1.77	0.151	1.17	0.705
Slightly negative rate - moving out	0.94	0.831	1.13	0.784	0.80	0.638
Balanced rate	0.79	0.434	1.33	0.550	1.30	0.568
Slightly positive rate - moving in	0.90	0.717	n.a.		0.99	0.982
Positive rate - moving in	0.78	0.426	1.49	0.367	0.55	0.299
Strong positive rate -moving in	0.76	0.371	1.78	0.305	0.55	0.300
Very strongly positive rate - moving in	0.81	0.488	2.19	0.130	0.73	0.585

Table 3.7. *Continued*

	General Population Refreshment (GP) n=1377		UB Refreshment, new regions (UBR) n=1134		UB Refreshment, new in the last year (UBN) n=614	
	Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
Under 30						
Up to 5%	(ref)		(ref)		(ref)	
5% - 10%	1.07	0.739	0.82	0.726	0.76	0.638
10% - 15%	1.05	0.820	2.01	0.146	2.25	0.111
15% - 20%	1.33	0.175	1.30	0.573	1.46	0.469
20% - 25%	0.90	0.626	1.12	0.812	1.49	0.393
25% - 30%	1.49#	0.088	0.57	0.284	1.29	0.576
30% - 35%	1.32	0.220	0.67	0.470	1.59	0.349
35% - 40%	1.23	0.441	1.24	0.664	1.85	0.247
40% - 50%	1.06	0.788	0.47	0.138	0.89	0.807
Over 50%	0.94	0.817	1.23	0.650	1.23	0.654
Foreign						
No foreigners	(ref)		(ref)		(ref)	
Extremely low proportion	0.82	0.345	1.10	0.857	1.80	0.182
Very low	1.16	0.535	1.62	0.345	n.a.	
Well below average	1.27	0.269	n.a.		1.00	0.993
Below average	1.08	0.752	0.74	0.548	1.16	0.757
Slightly below average	1.12	0.656	1.21	0.705	1.38	0.501
Average	0.83	0.424	0.52	0.223	1.48	0.401
Above average	0.82	0.439	1.06	0.905	1.22	0.662
Highest proportion	0.82	0.445	0.72	0.524	1.40	0.461
Family type						
Mostly single person households	(ref)		(ref)		(ref)	
Well above average proportion of single person households	1.37	0.294	0.49#	0.095	0.76	0.538
Above average proportion of single person households	0.90	0.726	0.77	0.528	1.14	0.760
Slightly higher than average proportion of single person households	1.01	0.968	0.78	0.565	0.81	0.632
Mixed family structure	0.94	0.847	0.59	0.255	1.22	0.648
Slightly higher than average proportion of families with children	1.11	0.753	n.a.		0.42	0.108
Above average proportion of families with children	1.04	0.907	0.49	0.127	0.38#	0.095
Well above average proportion of families with children	1.05	0.880	2.12	0.149	0.78	0.657
Almost exclusively families with children	0.67	0.236	1.52	0.514	1.66	0.427

Table 3.7. *Continued*

	General Population Refreshment (GP) n=1377		UB Refreshment, new regions (UBR) n=1134		UB Refreshment, new in the last year (UBN) n=614	
	Odds ratio	p value	Odds ratio	p value	Odds ratio	p value
Status						
Lowest social status	(ref)		(ref)		(ref)	
Very low status	1.36	0.239	1.85#	0.067	0.93	0.851
Well below average status	1.93**	0.007	1.32	0.457	0.92	0.829
Below average status	1.69*	0.043	1.49	0.332	1.59	0.294
Slightly below average status	2.03**	0.007	0.88	0.801	0.76	0.574
Average status	2.58**	0.000	1.67	0.247	1.29	0.561
Slightly above average status	2.96**	0.000	2.35#	0.059	0.71	0.478
Above average status	3.34**	0.000	1.95	0.220	1.95	0.185
Highest social status	5.80**	0.000	4.17*	0.014	5.92**	0.001
Cut point between low & med income (se)	0.37	(0.451)	3.02	(0.905)	1.26	(0.810)
Cut point between med & high income (se)	2.11	(0.454)	4.88	(0.928)	3.17	(0.827)

# p< 0.10, * p< 0.05, ** p< 0.01

3.4.3 Cross validations for UB

The predictive power of each type of auxiliary data observed above is illustrated with cross validation. First, the results for UB are presented. The mean of the squared differences between the survey values for UB (0, 1) and the predicted values (this mean is the "score" in figure 3.1) is plotted for each validation subsample of each of the three PASS samples. The detailed results showing the differences between the means and the significance of this difference from the paired t-tests are shown in appendix 3F (table F1) for all UB cross validation subsamples, across all three PASS samples.

For the general population (GP) refreshment sample[15], shown in figure 3.2, the plot and tests of significance show that better predictions of UB result when using the interviewer observations only or both sources, depending on the subsample. For example, in the second subsample, the mean deviation between the predictions and the survey values using the interviewer observations only is significantly smaller than the deviation using only the Microm indicators ($t(212) = -2.36$, $p = 0.019$). In this subsample, using both data sources has a marginally smaller deviation than using Microm only ($t(212) = -1.79$, $p = 0.074$). Using both data sources also performs better than only Microm data in the third subsample ($t(258) =$

[15] Attempts were made to improve the models presented by manipulating and dropping covariates to develop the "best" *unique* model for each sample. However, tests of fit and discrimination showed that the revised models could not improve upon the models reported. More importantly, the cross validations were very similar in pattern and significance, with only noticeable differences between some of the subsamples for GP, showing that the significance of the "observation-only" model over the "both" model is true less often (but the nonsignificant income observation was not a part of the "best" model for UB). Nonetheless, the Microm-only models consistently perform worse than the other two, as reported above.

-2.92, p = 0.004). Across the five subsamples, Microm data contain less accurate information about UB than the observations, but there is no consistent difference between using either the observations only or both the observations and Microm to obtain the best predictions.

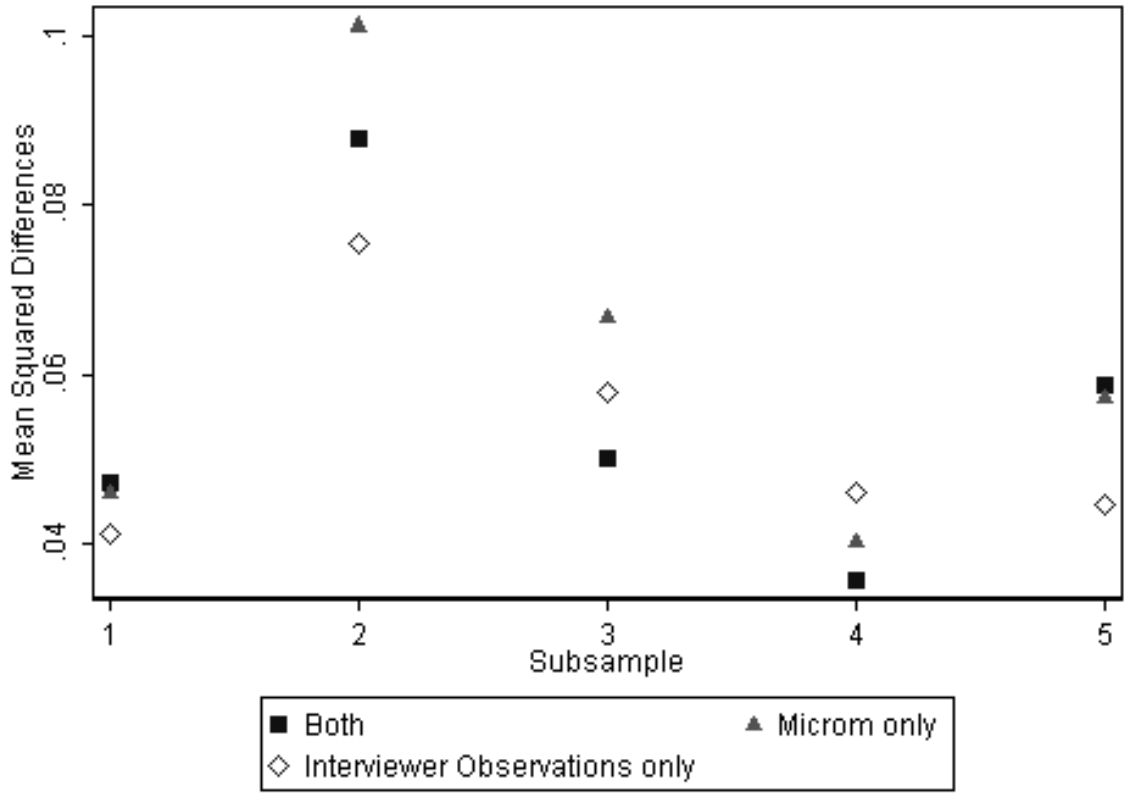


Figure 3.2. Plot of the Mean of the Squared Differences between the Predicted UB Value and the Survey Value for the Three Models Tested, Shown for Each Subsample of the GP PASS Sample

Figure 3.3 shows the results of the cross validation for the unemployment benefit *regional* (UBR) refreshment sample. In this PASS sample, it is evident that the interviewer observations are consistently better at predicting UB than either of the other models, and the model using Microm data only is the least predictive. All differences in subsample three are significant at the $\alpha = 0.05$ level. Subsample one also finds that both the observation-only model and the model using both sets of auxiliary data result in a significantly smaller deviation from the survey value than using the Microm data ($t(235) = -3.72$, $p = 0.0002$ and $t(235) = -2.88$, $p = 0.004$ respectively). If tested at the $\alpha = 0.10$ level, the observation-only model is also significantly better than using both sets of auxiliary data. At this level, additional significant differences are present in subsamples four and five, revealing similar conclusions.

Finally, analysis of the refreshment sample of those *new* to unemployment benefits (UBN) echoes the results from the UBR sample. As shown in figure 3.4, all subsamples find that the models using Microm data only are the least predictive while those using observations only are the most predictive of the survey value. Comparing the models using observations only and Microm only, the difference is significant at the $\alpha = 0.05$ level for four of the five subsamples. At this level, in three of the five subsamples, the observations are significantly more predictive than using both sets of auxiliary data.

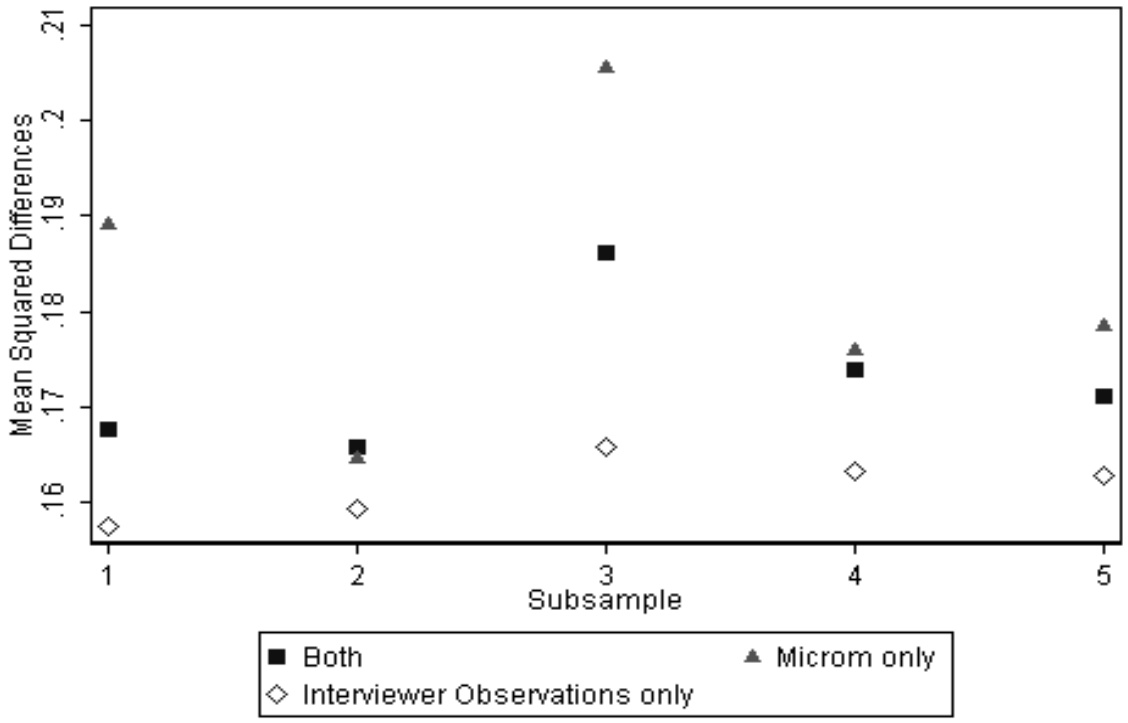


Figure 3.3. Plot of the Mean of the Squared Differences between the Predicted UB Value and the Survey Value for the Three Models Tested, Shown for Each Subsample of the UBR PASS sample

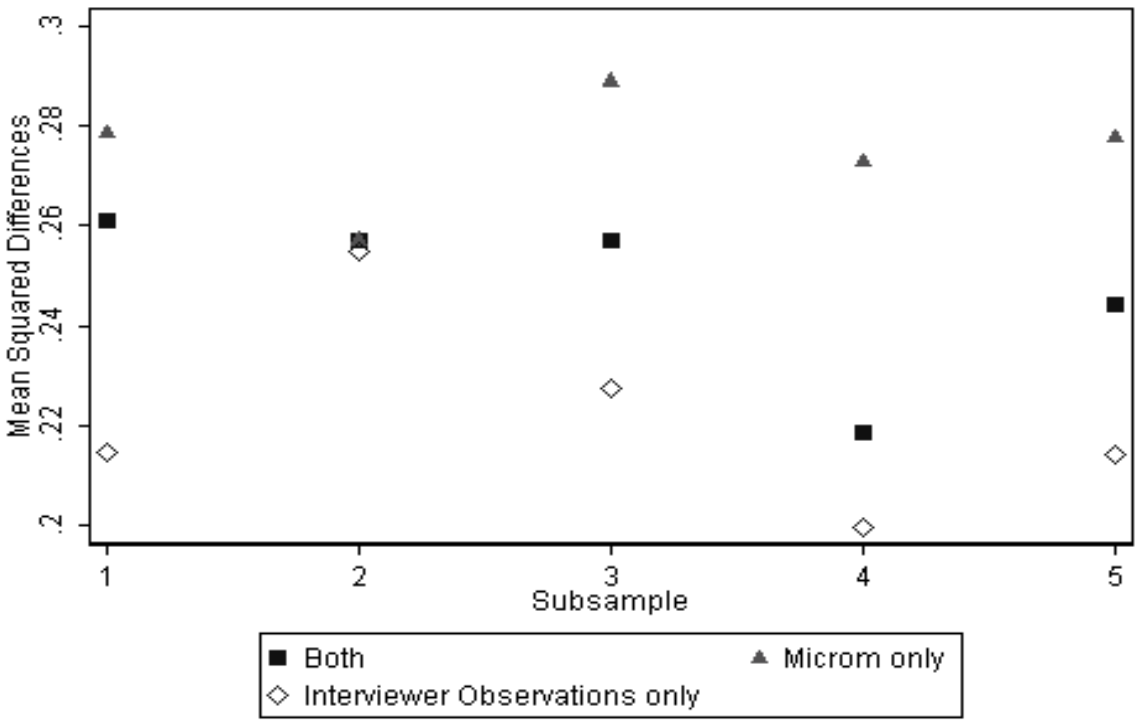


Figure 3.4. Plot of the Mean of the Squared Differences between the Predicted UB Value and the Survey Value for the Three Models Tested, Shown for Each Subsample of the UBN PASS sample

For UB, there seems to be agreement between the UBR and UBN samples that the observations are significantly more predictive than Microm data, and often also significantly better than the model using both sets of auxiliary data. This indicates that error in the Microm data may hurt the strength of the model using both sources. The GP sample is less conclusive but does show that the Microm-only models are the least predictive of UB. For GP, it is not clear whether using both sets of auxiliary data or only the observations is better. Of the three PASS samples, the GP model predictions had the smallest deviations from the survey value, resulting in most squared differences being less than 0.1. With such small values, it is difficult to find significant differences in the t-tests, given the sample size.

3.4.4 Cross validations for income

The results for the prediction of income in the GP sample, illustrated in figure 3.5, find significant differences at the $\alpha = 0.05$ level between using both data sources and Microm data only for all subsamples. For all subsamples but the fifth, the observations were also significantly better than the Microm data at predicting income. However, there were no significant differences in the mean deviations from the survey value when comparing the observations only to using both data sources.

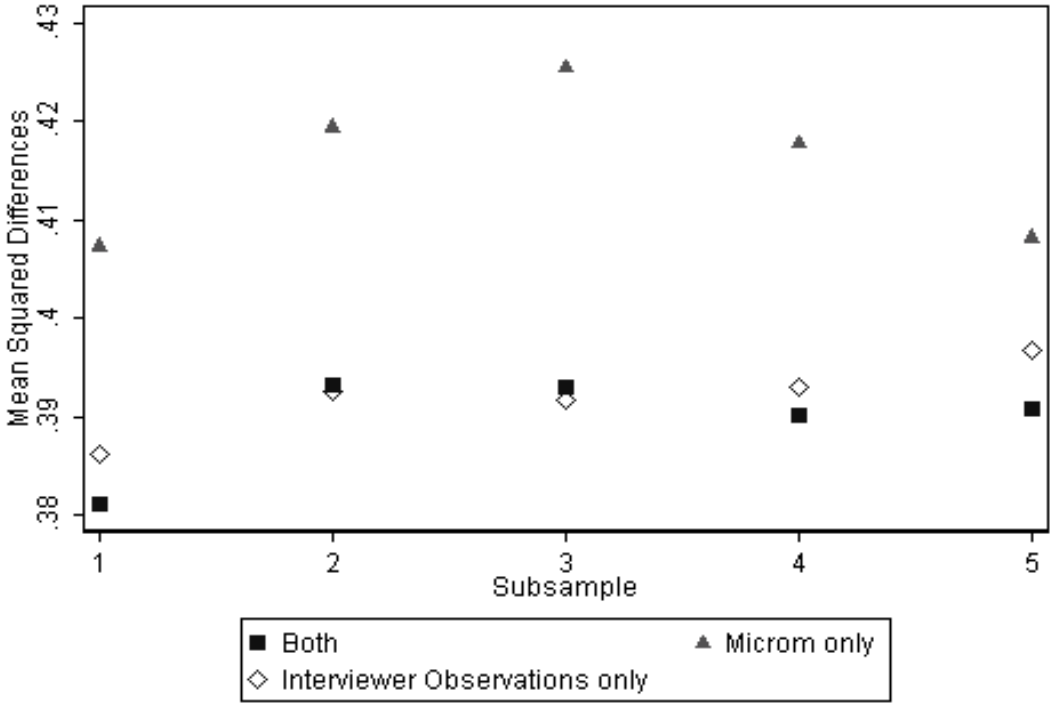


Figure 3.5. Plot of the Mean of the Squared Differences between the Predicted Probability of the True Income Category According to the Survey Data and 1, for the Three Models Tested, Shown for Each Subsample of the GP PASS Sample

The UBR and UBN samples do find that the observations-only model is more predictive than using both data sources in four of the subsamples (subsamples two, three, and five for UBR; subsample four for UBN). The results for these subsamples also show that the observations are significantly better than using Microm data only[16]. In these PASS samples, there is no significant difference between using the Microm data only and both data sources (see figures

[16] In subsample five of the UBR sample, the difference is marginal ($p = 0.087$).

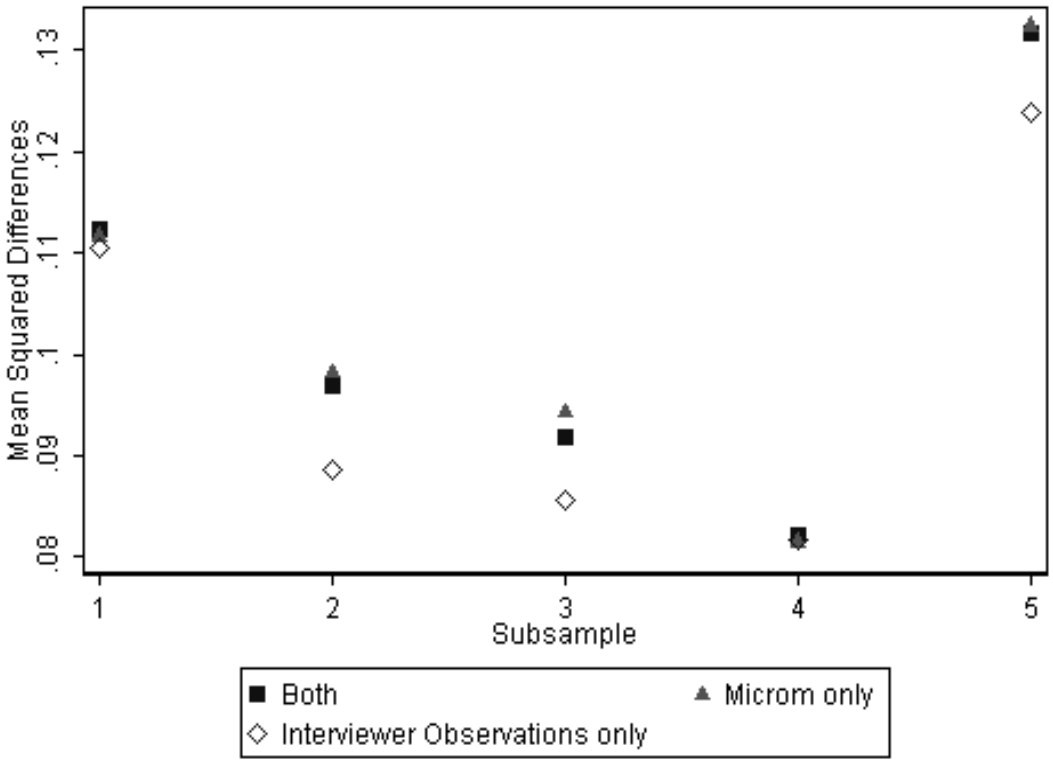


Figure 3.6. Plot of the Mean of the Squared Differences between the Predicted Probability of the True Income Category According to the Survey Data and 1, for the Three Models Tested, Shown for Each Subsample of the UBR PASS sample

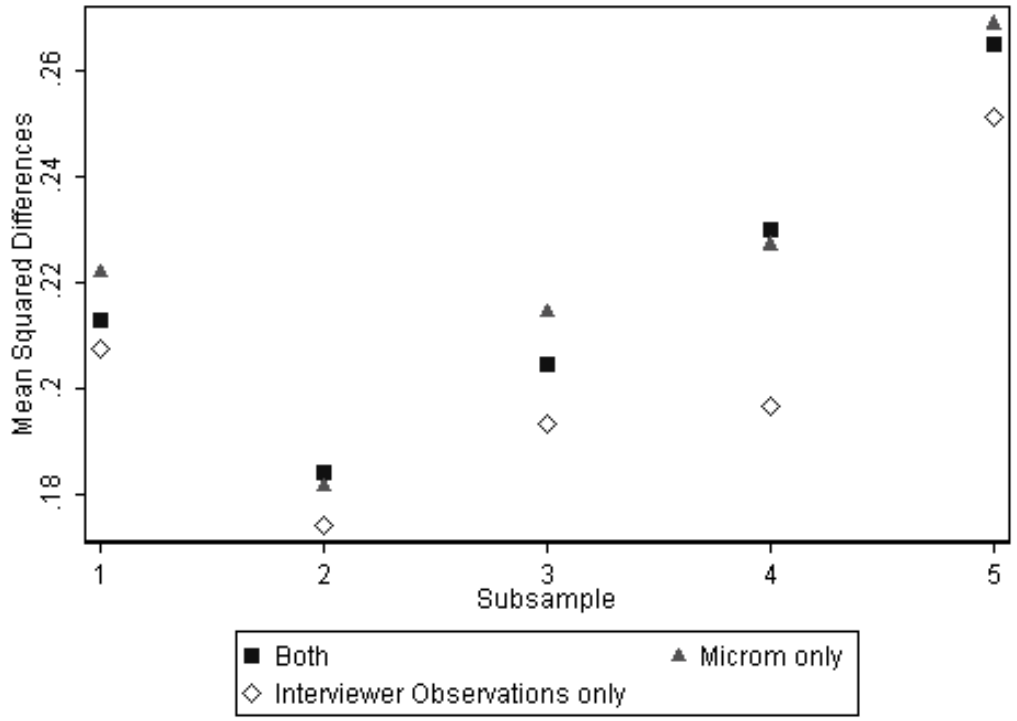


Figure 3.7. Plot of the Mean of the Squared Differences between the Predicted Probability of the True Income Category According to the Survey Data and 1, for the Three Models Tested, Shown for Each Subsample of the UBN PASS sample

3.6 and 3.7 for an illustration). As with UB, the detailed data to complement the cross validation plots of income for all three PASS samples can be found in appendix 3F (table F2).

When determining the best single data source to predict income, the analysis shows that the observations contain more accurate information than the Microm data. There is agreement between the UBR and UBN samples that the observations are significantly more predictive than using either the Microm data or both data sources. Analysis of the GP sample finds that using both data sources consistently performs as well as using interviewer observations only, although one can safely state that the Microm-only models are the least predictive.

3.4.5 Cost

For survey agencies that collect their own data, methods of calculating the cost of interviewer observations can vary depending on the pay structure of the interviewers and, if absorbed into the time and payment for other duties, can be difficult to extract. The PASS data are not collected in-house and therefore, the fees of the data collection agency provide an exact cost. Collecting the interviewer observations for the refreshment samples in wave five of PASS cost 8095 Euros, before taxes[17]. The cost of the Microm data was comparable, at 8177 Euros (before tax) for all variables for 30,000 households[18]. Therefore, an argument cannot be made for choosing one auxiliary data source over another based on expense alone. In addition, the purchase of either type of auxiliary data comprises less than one percent of the annual PASS budget, making the individual costs even less of an issue.

3.5 Discussion

This analysis used data from a German economic household survey (PASS) to answer the question of which type of auxiliary data—paradata in the form of interviewer observations or commercial data from the consumer marketing organization microm—is more predictive of the self-reported values from the survey and therefore a better investment of the survey budget for addressing nonresponse bias. The analysis used the two types of auxiliary data to predict two important study outcomes from the survey data: whether anyone in the household is on unemployment benefits (UB) and the categorized level of income. A simple comparison between the interviewer observations and the survey values showed moderate to poor agreement, and tests of independence between the Microm indicators and the survey outcomes were significant with acceptable to low strengths of association. The multivariate analysis, using logistic and ordered logit models, showed that the interviewer observations, particularly the observation designed to match the survey question analyzed, appear to be better predictors than the Microm data. Cross validation supported this conclusion that the observations contain more accurate information about the survey value, especially in the refreshment samples drawn from benefit recipient registries (UBR and UBN). In these samples, the observations were often more predictive than using both data sources. However, in the general population, using both auxiliary data sources also performed well. Across all three PASS refreshment samples for both survey outcomes, the Microm data were consistently the least predictive.

Although I conclude that using only Microm data is not the best predictor of the survey outcomes for any of the samples, the Microm data do contribute to the strength of the models

[17] This amount is inflated by the extra costs for executing the interviewer observation experiment and training interviewers who had not previously collected such information.

[18] Again, this cost is inflated because only a fraction of those households and a handful of relevant Microm variables were used in this analysis.

with both sources. While these models perform well in the GP samples, that errors in the Microm data may hurt the predictive power of the model using both sources for the UBR and UBN samples. This difference in the contribution of the Microm indicators is not surprising since these data are developed for use across the general population and are not intended to differentiate households with similar income, social status, and mobility within a subpopulation. This raises the issue of whether commercial data may not be the best choice for special subpopulations. Capturing data about such subpopulations appears to be better served by collecting interviewer observations.

One limitation of the analysis is that survey data represent the "true value." There may be measurement error in the self-reported survey data (Sakshaug and Kreuter 2012 identified underreporting of UB and overreporting of income in the first wave of PASS). Also, results are limited to respondents and it cannot be tested whether the accuracy of the auxiliary data for nonrespondents differs from the accuracy for respondents. Given differential measurement error on interviewer observations between noncontacts, refusals, and cooperative cases (Sinibaldi et al. 2013), the performance of interviewer observations in the analysis reported above may be somewhat diminished if all sample cases are analyzed.

Another possible limitation stems from interviewer behavior. Despite instructions, interviewers may have recorded the observations after the first visit, or even after the interview, improving their predictive power in the analysis. Given the moderate to poor agreement between the observations and survey data, however, the incidence of this behavior is estimated to be low. Additionally, a sensitivity analysis that dropped all cases where observations were not entered electronically at least one day before the household interview (dropping 54% for GP, 62% for UBR, and 67% for UBN of the final case count in table 3.2) and repeated the cross validation analysis found similarly strong, if not stronger, differences between using observations only and Microm data only or both to predict UB, even with degrees of freedom below 100 cases for all subsamples. The sensitivity analysis for income shows patterns and relationships consistent with the results presented but fewer differences are significant. This supplemental analysis not only addresses concerns about bias to the results due to interviewer deviations from protocol but also somewhat eases concerns about differences in the accuracy of observations between respondents and nonrespondents, since the observations used in the sensitivity analysis were certainly recorded while the cases were still "nonrespondents".

Based on their ability to predict the survey values in the data, the analysis finds that for nonresponse adjustment, interviewer observations are a better choice than commercial data, and their predictive power is especially notable for subpopulations with characteristics specifically captured by the observations. However, the relationship between propensity to respond and each auxiliary data source (a second criterion for good nonresponse adjustment) was not addressed. Assuming that survey topic is correlated with propensity to respond (Groves et al. 2000), one would expect the observations (which capture the survey topic) to be highly correlated and therefore, reinforce the conclusions that they are the best choice for reducing nonresponse bias. But efficient adjustment should also minimize variance. Although interviewer observations share more accurate information with the key survey outcomes than Microm does, they cannot be considered accurate. If the correlations between the observations and the true value are low due to error in the observations, then the resulting increase in variance of estimates computed using sampling weights that have

been adjusted for nonresponse using interviewer observations will inhibit effective bias correction (West 2013b).

In addition to quality, another major consideration in purchasing auxiliary data is cost. The costs of collecting interviewer observations or purchasing commercial data for PASS were comparable and constituted less than one percent of the annual budget. Therefore, compared to the expenditures of a large government survey, the cost of purchasing either source of auxiliary data, or funding both, is likely to be minimal. An example of this decision in a large survey comes from the University of Michigan. An internal analysis calculated the potential reduction in the cost per interview on the National Survey of Family Growth (NSFG) if commercial data were available. The report concluded that "if the use of the [commercial vendor] data helped NSFG complete 100 more interviews a year for the same cost, half of the [commercial vendor] cost would be justified" (Hubbard and Lepkowksi 2009). These purchase decisions may not be as straightforward for projects with small budgets that cannot as easily absorb the expense.

A relevant component of cost is the "2-for-1" benefit of auxiliary data sources. Commercial data have many indicators and can be used for more than one survey, or wave of a survey, if all necessary geographic areas are purchased. In addition, commercial data can be appended to the sampling frame to assist in sample selection and they can be used in nonresponse weighting. Interviewer observations cannot offer these advantages since they are collected after the sample is drawn and only for households selected for a particular survey and for a limited number of characteristics. However, if the results of this analysis are broadly applicable, interviewer observations designed to target a specific characteristic of interest are the higher quality source. Given the low cost of purchasing both types of data, it would be wise to collect observations *and* purchase commercial data, especially when using these data for nonresponse adjustment of the general population.

The question remains, to what extent can interviewer observations correct for nonresponse bias? Previous work is limited but shows that their ability to improve adjustments is minimal (Kreuter et al. 2010b; West et al. forthcoming). That work, research on the quality of interviewer observations (Sinibaldi et al. 2013; West 2013a), and the conclusions stated above all note that the potential of interviewer observations is likely hampered by shortcomings in their quality, and all encourage efforts to improve their accuracy. Once this is achieved, the value of interviewer observations for nonresponse adjustment can be more fairly assessed.

Chapter 4: Improving Response Propensity Models with Interviewer Observations

4.1 Introduction

Although response propensity models are traditionally used for developing nonresponse weights after the close of data collection (Little 1986), these models which predict the likelihood of a unit to cooperate to the survey request are increasingly being used during data collection as well. Propensity modeling is an integral part of the fieldwork monitoring in responsive survey designs (Groves and Heeringa 2006) and adaptive survey designs (Schouten et al. 2013). Applications within the responsive survey design framework involve the generation of propensity models at regular intervals during data collection to: direct face-to-face interviewers to work the cases most likely to cooperate (Wagner et al. 2012), choose the best time for the call scheduler to make the next telephone call (Wagner 2013), and calculate the R-indicator (Schouten et al. 2009) to determine when efforts are no longer improving the representativeness of the sample. Outside of responsive survey design, propensity models are generated during fieldwork to more accurately evaluate interviewers' performance, such as in the calculation of the propensity adjusted interviewer performance (PAIP) indicator (West and Groves 2013).

In order for these field techniques to be successful, the propensity models must have sufficient predictive power for the purpose. However, the model fit of propensity models, as designated by the pseudo R^2 values, is typically poor (e.g., pseudo R^2= 0.032-0.077, Olson et al. 2012; pseudo R^2= 0.022, Olson and Groves 2012; pseudo R^2= 0.074 - 0.337, West and Groves 2013). To obtain high values of model fit, the covariates should be of good quality for both respondents and nonrespondents and highly correlated with cooperation. Several analyses using propensity models with the aim of improving nonresponse adjustment, have concluded that the paradata and auxiliary data currently available on respondents and nonrespondents are not sufficiently correlated with cooperation and key survey outcomes (Peytcheva and Groves 2009; Kreuter et al. 2010b; Olson and Groves 2012) or of satisfactory quality (Biemer and Peytchev 2012; Biemer et al. 2013; Sinibaldi et al. *forthcoming*). Within this general call for better paradata, West and Groves (2013) have specifically argued for new or better quality paradata to improve predictions of response (i.e. response propensity models), especially for telephone studies.

Given this need for new sources of paradata, available on both respondents and nonrespondents, to better model response propensity, this analysis examines a new type of paradata -- call-level interviewer ratings of response likelihood (see Eckman et al. 2013 for a descriptive analysis of these ratings)-- to determine if these data can improve the predictive power of propensity models. This analysis will use typically available call record and interviewer paradata to develop a "classic" discrete time response propensity model used for responsive survey design. Then, the interviewer ratings of response likelihood will be added to this model to create a new version of the model. Several tests of fit and discrimination will determine if the classic model is significantly improved by the inclusion of the new paradata. An additional test will involve using the two versions of the model, along with the ratings on their own, to estimate "daily" response propensities in the context of responsive survey design. The analysis will compare the ability of each version of these propensity models to predict the cases most likely to cooperate on the next contact at several time points during

the data collection. The conclusions will pool the results from the tests of model improvement and the performance of the models in a responsive survey design context to determine the value of the new interviewer ratings for estimating response propensity, particularly during fieldwork.

4.2 Data

The data analyzed are from an experimental CATI[19] survey conducted in Germany from October 29 to December 14, 2012, designed to study methodological determinants of measurement error. This was the second wave of the data collection and the survey included topics about employment status, income, and socio-demographics. In wave 1, a sample of 12,400 adults who are or were previously employed was selected from three strata from German administrative databases. In total, 2,400 interviews were completed in the first wave, yielding a response rate of 19.4% (AAPOR RR1). Of these, 87% (2,085 people) agreed to be contacted again. Only 1,324 of these responded in wave 2, yielding a response rate of 63.5% (AAPOR RR1). The analysis primarily uses call records from wave 2, since the interviewer ratings of likelihood are collected for this wave. Analyses conducted during model development also included data from the wave 1 call records and a survey of the CATI interviewers but these data were not relevant for the final models and results.

4.2.1 Likelihood ratings

At the end of each contact attempt, interviewers rated the likelihood of the selected target person at that household to complete the survey on a later call, using a scale from zero to 100. The text of this question, translated from German, was:

> How likely is it that this case will complete the interview at a later contact attempt? Please give the probability in percent, from 0 to 100.

All 16,318 calls were rated except those resulting in an interview, handled entirely by the autodialer, and very hard refusals. Interviewers were not able to see the ratings assigned to the same case by other interviewers on previous calls, if any, and could not skip the question or respond “don’t know.” Cases were assigned to any available interviewer and there are no refusal conversion specialists. Therefore, that the assignment of cases to interviewers is assumed to be random. All CATI interviewers had worked over a year at the data collection agency and would have made these likelihood observations on a prior study.

4.2.2 Cleaning of contact records

All calls that did not result in contact were dropped from the analysis, leaving 5,049 contacts. This was done for two reasons. First, there is a precedence to analyze contacts only, such as in the calculation of the PAIP (West and Groves 2013). Second, the interviewer ratings are most applicable to cases with contact since a call with no contact provides no information on which the interviewer can make a sensible rating. Using contact calls only affects the interpretation of the results because the analysis predicts cooperation, assuming contact. Therefore, for the results to be useful, the data collection agency has to do their part and make contact.

Two additional changes were made to data. First, due to a small number of cases with 14 -18 contacts which affected model performance, the data were censored to exclude all contacts

[19] Computer Assisted Telephone Interviewing

greater than 13 (removing 15 contacts and two interviews). Second, since the likelihood ratings were recorded at the end of the contact and refer to future call attempts, the ratings were lagged forward so that they are associated with the next call with contact (see table 4.1 for an illustration). This procedure provided ratings for the call when the interview was taken for those cases that did not cooperate on the first contact. However, lagging the ratings resulted in missing ratings for the first contact of all cases in the analysis. The likelihood rating for the first contact was imputed a couple different ways, as explained in appendix 4A, but ultimately the first contact was dropped. This reduced the number of contacts in the analysis to 3091 (and removed 505 interviews that were completed on the first contact and 143 cases with only one contact). The final analysis dataset had 817 interviews across 1295 cases and 22 interviewers. The number of cases available for each contact number ranged from 9 to 1295 (see appendix 4B). Each interviewer is associated with 2 to 333 contacts (mean=140.5; sd=90.4)[20].

Table 4.1. Snapshot of the Dataset Showing the Forward Lagged Likelihood Rating and the Deletion of the First Contact

Case ID	Contact Number	Cooperated	Likelihood Rating	Lagged Rating	Categorized Lagged Rating	Average Lagged Rating
1	2	0	30	20	3	20
2	2	0	10	80	10	80
2	3	0	1	10	2	45
4	2	1	.	60	8	60
5	2	0	75	70	9	70
5	3	0	50	75	9	73
5	4	0	99	50	6	65
5	5	0	100	99	12	74
5	6	0	50	100	12	79
5	7	0	75	50	6	74
5	8	0	50	75	9	74
5	9	0	75	50	6	71
5	10	0	50	75	9	72
5	11	0	60	50	6	69
5	12	0	55	60	8	69
5	13	0	0	55	7	67

4.2.3 Manipulation of the likelihood ratings

The distribution of the likelihood ratings shows significant rounding, indicating uncertainty in the interviewers' predictions (Tourangeau et al. 2000). The interviewers use mostly the tens categories, especially below the rating of 50 (see figure 4.1 in the results section). Above 50, the categories ending in five are used more often than below 50. An early attempt to categorize the ratings used a 2.5 point interval around the ratings ending in five or zero, resulting in 21 categories. This scale performed much the same as the continuous rating, with no noticeable change to the parameters or fit of the models. An alternate categorization used 12 categories, collapsing values from 0 to 9 within each decile as a single category for

[20] The interviewer with two contacts conducted interviews on both calls. Therefore, the interviewer did not provide ratings but since these not the first contacts with the cases, the lagged rating from the previous contact is applied to the contacts.

all but the 50s and the 90s. Since the rating of 50 was used for 21% of the ratings, it was given its own category and consequently, the next category higher contained ratings of 51-59 (see case 5, contacts 7 and 13 in table 4.1). The ratings of 90 and higher were divided into 90-95 and 96-100 out of theoretical interest, to see if the highest ratings were most predictive of cooperation.

In addition to the categorized lagged rating, a second variable using the likelihood ratings was created and tested in the models. This variable averages the ratings assigned to a case prior to the current call. Since the dataset excludes the first contact, the average of the rating at the second contact is the rating made at the end of contact 1. The average rating at the third contact is the average of the ratings made at contacts 1 and 2, etc. (see table 4.1).

4.2.4 Interviewer data

During the data collection, a voluntary survey of interviewers was conducted. The survey captured demographic information, work experience, satisfaction, and personality characteristics. The analysis includes only variables that are likely to be available to a survey researcher and therefore limited the pool of variables to characteristics commonly found in employee records. Some of these variables suffered from significant item missing data, leading to the decision to use only three variables: months/years of experience at the data collection agency, hours worked per week at the agency (currently), and age (which directly corresponds with student status, since all interviewers born at or after 1980 are students).

All variables used in the modeling are explained in appendix 4C.

4.3 Methods

4.3.1 Overview

The analysis begins with a descriptive exploration of the likelihood ratings, graphically displaying the distribution of the ratings over all contacts in the analysis, by contact outcome, and by interviewer. These analyses provide some information as to how the interviewers use the ratings. Following this, a bivariate analysis explores the relationship between the ratings and cooperation at the next contact. The relationship is characterized both graphically and statistically, using a Chi-square test.

The multivariate analysis is divided into two parts, which are explained here briefly and in more detail below. First, response propensity models are developed using all contacts from the complete wave of data collection. The models are labeled as “Classic” to indicate a propensity model that uses available call record and interviewer data, and “Classic+” to indicate a Classic model that also includes the interviewers’ ratings of likelihood to respond. For comparison, a response propensity model with only the interviewers’ ratings, called the “Ratings-only” model, is also presented. The ability of the three versions of the models to predict cooperation is compared by analyzing fit and discrimination statistics. In the second part of the analysis, the three models are applied at the close of several dates during the data collection to simulate the daily propensity modeling that would be run during a live responsive survey design. The performance of the models in this application is assessed by comparing the percent of cases that are accurately predicted to cooperate on the next contact by each model for each date examined.

For all parts of the analysis, the response propensity models are discrete time logistic hazard models (Singer and Willett 1993; Durrant et al. 2013) using contact number as the discrete time variable and predicting cooperation on the next contact. Hazard models (also called survival models) are favored over a propensity model that aggregates to the case because time variant information (i.e., the detailed call history) is included in the model. Hazard models are preferred for fieldwork monitoring when implementing responsive survey design (Wagner and Hubbard *forthcoming*) and it is this technique that this analysis aims to improve.

4.3.2 Equations and details of modeling strategy

The form of the discrete time logistic hazard model is as follows (based on Singer and Willett 1993):

$$\ln\left(\frac{h_{i,t}}{1-h_{i,t}}\right) = B_{0,t} + \boldsymbol{B}\boldsymbol{x}_{i,t} = B_{0,t} + B_1 x_{1i,t} + \cdots + B_P x_{Pi,t} \quad (1)$$

where the hazard $h_{i,t}$ is the conditional probability that respondent *i* will cooperate at contact number *t*, given that the case did not cooperate at the previous contact, *t-1*. $B_{0i,t}$ is an intercept term that applies to all individuals at time *t*, $x_{i,t}$ is a vector of values of the covariates for respondent *i*, and $\boldsymbol{B}$ is the vector of the corresponding regression parameters. The covariates represented by $x_{i,t}$ are both time varying and time invariant.

The nature of the hazard function and the data is that a case can only cooperate once. So, if a case cooperates at contact t, it is not modeled at *t*+1. Cases that have not cooperated at contact t will be referred to as "active" cases. To obtain the probabilities of cooperation for the cases that are active, an inverse logit transformation must be performed:

$$\hat{h}_{i,t} = \frac{exp\left(\hat{B}_{0,t} + \hat{\boldsymbol{B}}\boldsymbol{x}_{i,t}\right)}{1 + exp\left(\hat{B}_{0,t} + \hat{\boldsymbol{B}}\boldsymbol{x}_{i,t}\right)} \quad (2)$$

In the analysis, contact number is a discrete time with indicators for each contact number. Therefore, the above equations can be expanded to denote the indicator for each contact number as D_t. D_t=1 when the contact number = *t*; otherwise, D_t=0. B_0 is replaced with α for clarity and the intercept removed so that all contact numbers are modeled.

$$\ln\left(\frac{h_{i,t}}{1-h_{i,t}}\right) = [\alpha_1 D_{1i,t} + \alpha_2 D_{2i,t} + \cdots \alpha_T D_{Ti,t}] + B_1 x_{1i,t} + \cdots + B_P x_{Pi,t} \quad (3)$$

Finally, the interviewers may differ in their ability to gain cooperation (West and Olson 2010, O'Muircheartaigh and Campanelli 1999). Therefore, the analyses run the discrete time logistic hazard models accounting for the random effect of interviewers (*j*), thereby capturing the variance in cooperation attributed to interviewers (u_{0j}). See Appendix 4D for a discussion of the advantages and disadvantages of using random versus fixed effects.

$$\ln\left(\frac{h_{ij,t}}{1-h_{ij,t}}\right) = \boldsymbol{B}_{0,t} + \boldsymbol{B}\boldsymbol{x}_{ij,t} + \boldsymbol{u}_{0j} \quad (4)$$

4.3.3 Developing the models

The Classic propensity model was developed in a stepwise fashion, introducing sets of related covariates or single covariates at each step to evaluate their significance in the model and how the addition affects covariates already added in previous steps. The cut-off for retaining a covariate for further steps was $\alpha = 0.10$. The covariates explored came from the call record and interviewer data, as explained in the data section. Since the analysis uses hazard models, including variables describing the history of the case (such as whether the person refused on a prior contact) is important. Although the effect of contact number could be linear or some other form, in these data it was best represented using discrete dummy variables, as shown in the equations above.

Introducing the interviewer ratings to create the Classic+ model necessitated investigations to account for the differences in the way the interviewers use the rating scale (as seen in Eckman et al. 2013). With each case worked by more than one interviewer, different approaches to modeling the interviewer effect were applied: fixed effects for the lagged interviewer (corresponding the lagged rating), interactions between the lagged interviewers and the lagged ratings, and random coefficients for the ratings accounting for the effect of the lagged interviewers. These tests concluded that a fixed or random effect for the lagged interviewer was not necessary. Results are summarized in Appendix 4E.

Once each model (Classic, Ratings-only, and Classic+) was finalized, a comparison of the model fit and discrimination was conducted to evaluate if the Classic model was improved by the addition of the likelihood ratings. The following were examined: pseudo R^2 values, AIC values, ROC curves, and likelihood ratio tests.

Since noticeably improving the area under the ROC curve can be difficult if the standard model (in this case the Classic model) is already strongly predictive of cooperation (Ware 2006), a supplemental evaluation of the improvement in discrimination was conducted. This evaluation involved first estimating the probability of cooperation for each contact in the Classic and Classic+ models and classifying these probabilities into tertiles (high, medium, and low). The tertiles were cross-tabulated to assess how many cases were classified differently between the two versions of the model. Then, to evaluate whether the difference in classification was an improvement or not, the Net Reclassification Index (NRI) was calculated (Pencina et al. 2008). The NRI provides the net improvement in a model when additional variables, in this case the ratings, are added. The calculation sums the difference between the proportion of cases that moved from the incorrect group in the Classic model to the correct group in the Classic+ model and the proportion of cases that moved the opposite direction. This is calculated separately for respondents and nonrespondents (see appendix 4F for formulas). So, the NRI takes into account not only the improvement in discrimination of new model but also penalizes the improvement by accounting for cases that were incorrectly reclassified.

4.3.4 Responsive survey design daily models

To evaluate the ability of the likelihood ratings to improve models used for regular monitoring under responsive survey design, the propensity to cooperate was estimated for twelve selected dates of the data collection (see appendix 4G for information as to how the dates were selected) for each of the three models: Classic, Ratings-only, and Classic+. Since there are fewer contacts and contact numbers to model early in the field period, the “daily”

models required some customization. All models in November excluded the flag for a refusal on the prior contact, and all models required that the highest contact number(s) be dropped. In addition, the number of prior appointments and the likelihood ratings could not be used as categorical for the early part of the field period. For consistency, these were kept as continuous variables for all daily models.

To evaluate which model is most predictive of cooperation at the next contact, the predicted probability of cooperation is estimated for all contacts prior to or on the date of the daily monitoring. The distribution of the probabilities from the most recent contact with each case is then divided evenly into thirds and categorized into probability tertiles, representing high, medium and low probabilities of cooperation[21]. A descriptive analysis examines differences between the boundaries of each probability tertile across the models by daily monitoring date. In addition, the classification of cases into the high, medium or low categories on each day is compared for the Classic and Classic+ models to evaluate whether cases are classified differently when the ratings are included in the model. Finally, the percent of cases in each probability tertile that cooperate on the next contact is compared across models. These "success rates" are examined for each daily monitoring date and in aggregate and are most revealing of the differences in the predictive power of the models.

4.4 Results

4.4.1 Descriptive analyses

The distribution of the forward-lagged ratings over all contacts in the analysis is shown in figure 4.1. It is clear that the interviewers tended to use the ratings ending in zero and, especially above 50, ending in five. Therefore, once the ratings were recategorized for the analysis, the distribution is similar (see figure 4.2).

What is notable in both the raw and recategorized distributions are the spikes at 50 and 60. Eckman et al. (2013) also found frequent use of the 50 rating in their analysis of the same 100 point scale and concluded that the use of this rating indicates "don't know" (Fischhoff and Bruine de Bruin 1999). This is also a plausible deduction for these data; 49% of the general contacts are rated with a likelihood score of 50 (see figure 4.3). If assigning a rating of 50 means "don't know", then assigning a rating of 60 seems to indicate "not sure". Of all of the contacts that resulted in appointments, 25% were given a likelihood rating of 60. Other than the rating of 50, the percentage of appointments assigned other ratings is small (10% or less; see figure 4.3). A rating of 60 seems to indicate that the likelihood to participate is better than chance but interviewers are reluctant to make a stronger prediction.

[21] Tertiles are used in the daily response propensity models run by the National Survey of Family Growth to determine the highest propensity cases in a segment (email with James Wagner, October 23, 2013). The US Census Bureau is using the same categorization for their responsive design testing (Miller 2013).

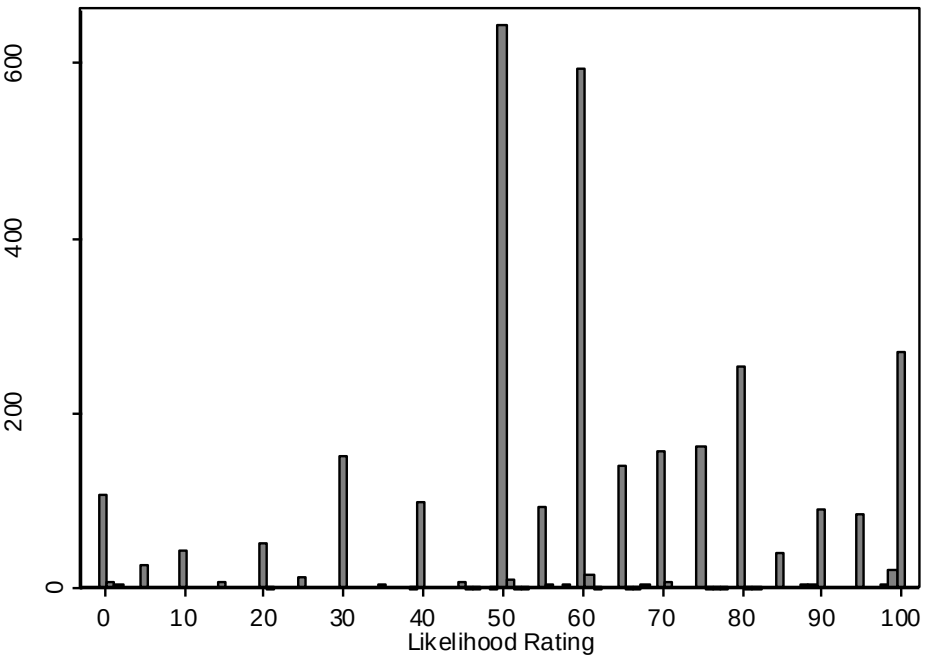


Figure 4.1. Distribution of Likelihood Ratings in the Analysis Before Categorization

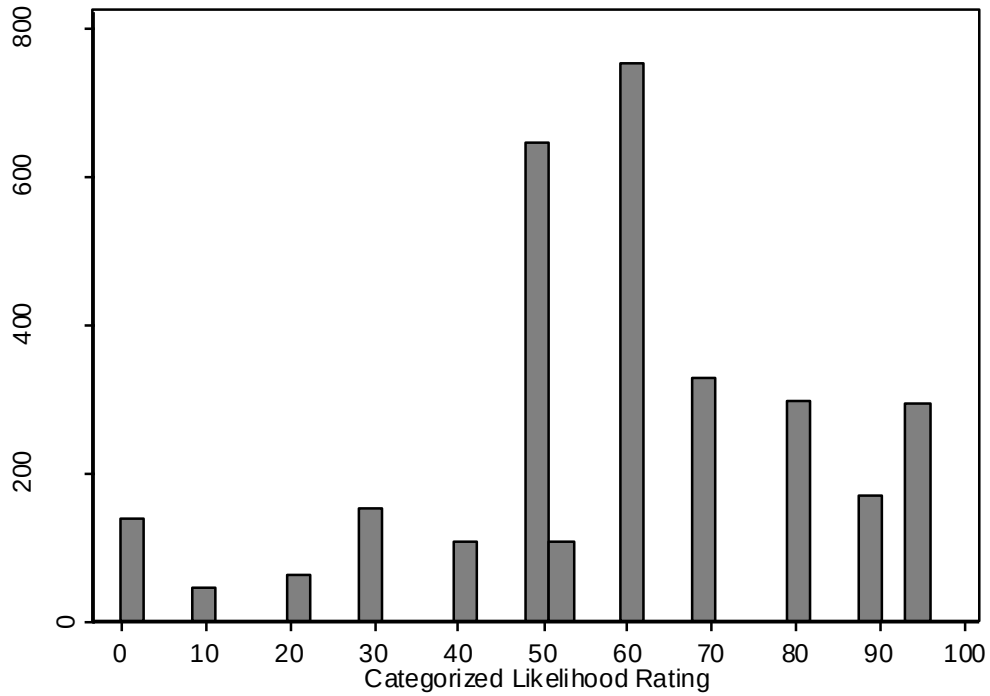


Figure 4.2. Distribution of Categorized Likelihood Ratings in the Analysis

For each possible contact outcome, figure 4.3 shows the distribution across the categorized likelihood ratings (not lagged forward) for each contact in the analysis. As noted earlier, this figure shows that the general contacts are mostly rated with a likelihood of 50 – 69 at the end of the call. The figure also shows that the contacts that result in refusals are mostly rated as 0 likelihood and the appointments mostly have high likelihood scores of 50 and above.

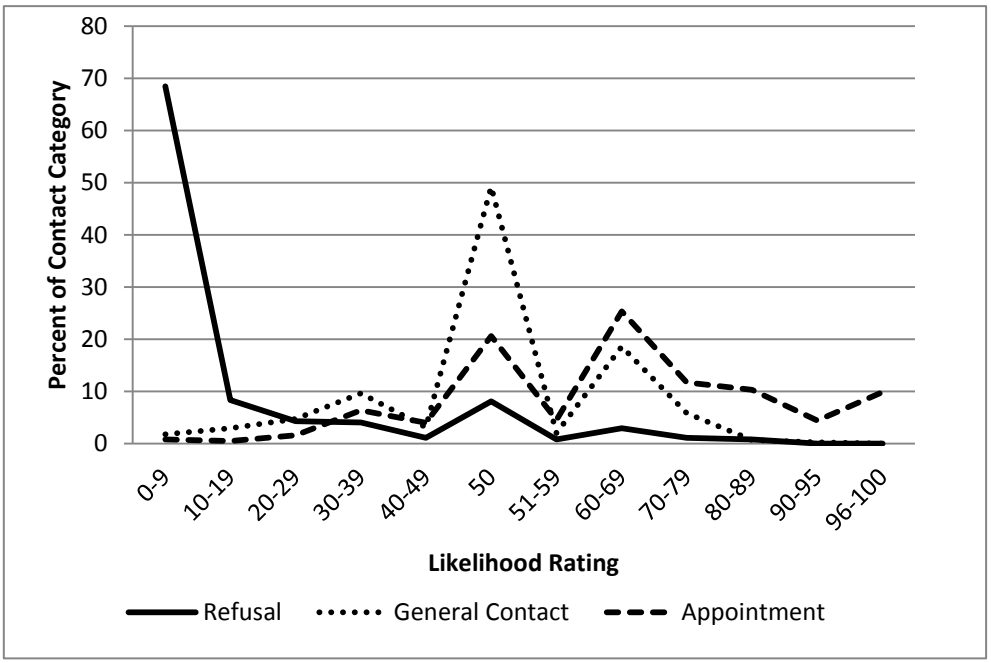


Figure 4.3. The Distribution of Categorized Likelihood Ratings Made at the End of Each Contact Call, for the Three Types of Contact Outcomes

Since the objective of this analysis is to understand if and to what extent the likelihood ratings predict cooperation, the relationship between the percent of contacts that resulted in an interview for each category of the forward lagged rating was examined graphically (see figure 4.4). The size of the marker in figure 4.4 indicates the number of contacts that received the categorized rating on the prior contact. The relationship between the rating and the outcome at the next call is clear; there is a notable linear trend between the increasing likelihood ratings and the percent of next contacts that result in an interview. However, the relationship is not 1:1 and therefore, a rating of 60 does not equate to a 60% completion rate. Eckman et al. (2013) also found that the interviewer ratings correlated with true completion likelihood and that the ratings could not be interpreted literally. A Chi-square test confirmed that the cooperation rates across the categorized lagged forward ratings are significantly different ($X^2(11)= 182$; $p=0.000$) and the strength of the association is moderate (Cramer's $V=0.24$).

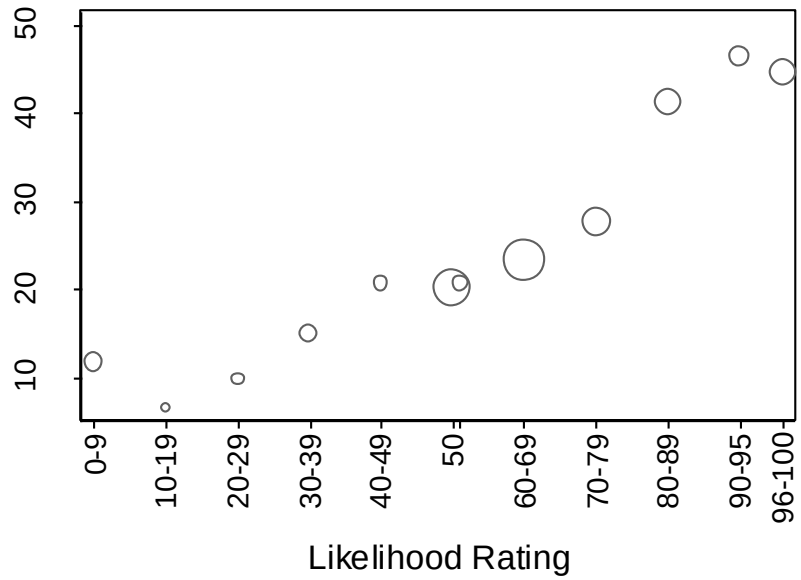
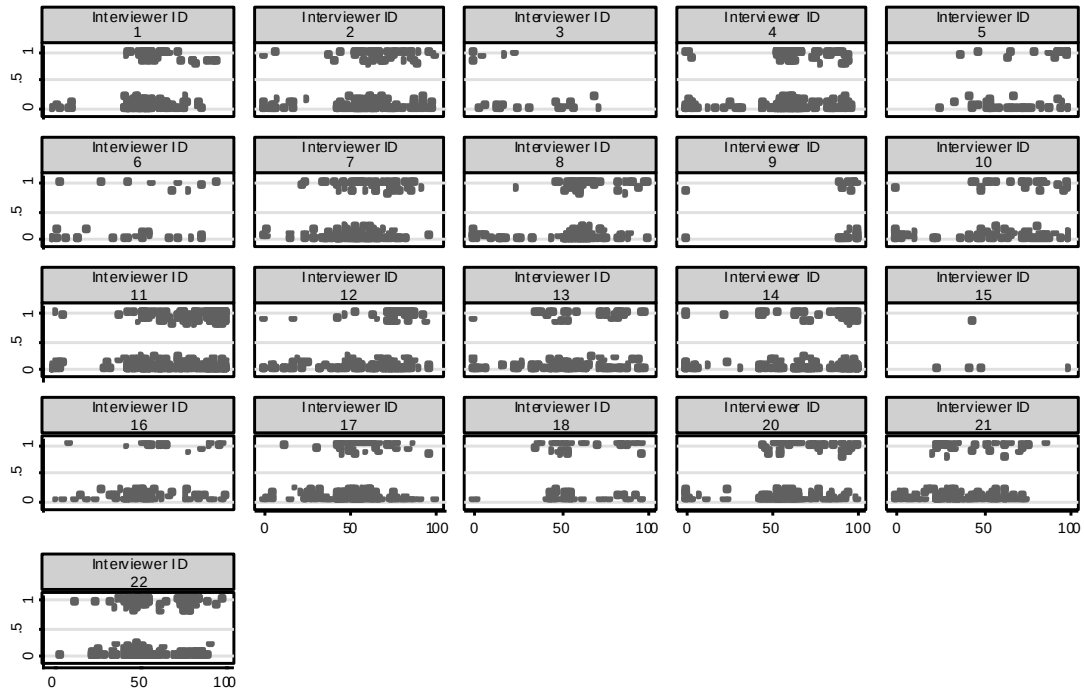


Figure 4.4. Percent of Each Categorized Likelihood Rating that Resulted in an Interview on the Next Contact; Showing Relative Case Bases for Each Category

The figures above show the likelihood scale in aggregate but it is interesting to understand if individual interviewers are using the scales differently. Tests interacting the lagged interviewer with the lagged rating found little evidence of a significant interviewer effect on the ratings' ability to predict cooperation (see apppendix 4E). Examining the distribution of ratings graphically, by cooperation on the next contact, shows that a few interviewers (e.g. interviewers 3 and 9) do use the range of ratings differently (see figure 4.5) but generally, the interviewers use the full scale with seemingly more ratings around and above 50. Since the interviewers were not deliberately assigned particular cases, the distributions of the ratings should be similar if the interviewers apply the scale in the same way. Although there are a few interviewers who use the scale differently, they do not introduce a statistically significant effect.



*Interviewer 19 not shown because he/she had 2 contacts, both interviews, and no forward lagged ratings.

Figure 4.5. Distribution of Likelihood Ratings Made by Each Interviewer by Cooperation or Not on the Next Contact; Points Jittered

4.4.2 Multivariate multilevel analyses

Table 4.2 displays the discrete time hazard models predicting the propensity to respond for the following models: (1) Empty, with only the contact numbers (2) Classic, without the likelihood ratings (3) Ratings-only and (4) Classic+ which includes the likelihood ratings. Note that although interviewer characteristics were tested in the models, these variables were not significant and did not reduce the random effect of interviewers. The contact numbers are not shown for parsimony but these coefficients can be found in appendix 4H.

Looking first at the intraclass correlation coefficient (ICC in table 4.2), the empty model shows a small but significant interviewer effect: two-percent of the variance in cooperation is attributed to interviewers. The effect is similar for the model with the ratings only but reduced when other covariates from the call records are included in the models (i.e., Classic and Classic+ models). The likelihood ratio tests comparing the standard logit model to the multilevel model are significant at α=0.01 for all models.

Table 4.2. Four Versions of the Discrete Time Hazard Propensity Models Predicting Cooperation, with Random Effects for Interviewers; Parameters shown as Odds Ratios with p-values

	Empty N = 3091		Classic N = 3091		Ratings-only N = 3091		Classic+ N = 3091	
Week			0.976 p=0.400				1.033 p=0.318	
Mobile phone			1.518 p=0.000	***			1.466 p=0.001	***
Wkday eve			0.705 p=0.001	***			0.689 p=0.000	***
Weekend			0.804 p=0.092				0.764 p=0.042	*
Num prev calls			0.974 p=0.024	*			0.967 p=0.006	**
Days since last contact			0.953 p=0.000	***			0.956 p=0.000	***
Refused previously			0.636 p=0.056				0.625 p=0.061	
Refused on prior contact			0.557 p=0.105				0.442 p=0.071	
No contact, prior call			0.686 p=0.000	***			0.679 p=0.000	***
1 previous appt			1.832 p=0.002	**			1.543 p=0.034	*
2-3 prev appts			3.107 p=0.000	***			2.431 p=0.002	**
4+ prev appts			4.550 p=0.000	***			3.557 p=0.001	**
Target person reached, prior contact			1.535 p=0.000	***			1.202 p=0.064	
Rating 10-19					0.485 p=0.273		0.226 p=0.036	*
Rating 20-29					0.714 p=0.511		0.303 p=0.040	*
Rating 30-39					1.166 p=0.677		0.318 p=0.017	*
Rating 40-49					1.504 p=0.280		0.396 p=0.060	
Rating 50					1.465 p=0.229		0.432 p=0.059	
Rating 51-59					1.364 p=0.427		0.373 p=0.048	*
Rating 60-69					1.524 p=0.198		0.375 p=0.030	*
Rating 70-79					1.737 p=0.126		0.464 p=0.105	
Rating 80-89					3.042 p=0.003	**	0.767 p=0.583	
Rating 90-95					3.136 p=0.006	**	0.825 p=0.710	
Rating 96-100					3.071 p=0.006	**	0.799 p=0.662	
Avg rating, start of call					1.010 p=0.024	*	1.005 p=0.305	
ICC (se)	0.021 (0.009)		0.010 (0.007)		0.018 (0.009)		0.013 (0.008)	

* p<0.05, ** p<0.01, *** p<0.001

When the ratings alone are used to predict cooperation on the next contact, only the ratings of 80 and above are significantly more predictive of cooperation than the 0-9 category. However, the direction of the parameter estimates, compared to the lowest rating, generally increases as expected and all contacts rated 30 or higher have odds ratios greater than one, indicating that these ratings are more likely to result in cooperation on the next contact compared to the lowest rating category (though not all estimates are significant). The increasing trend in the values of the odds ratios is consistent with the pattern shown in figure 4.4, as the odds ratios are less than one for the 10-19 or 20-29 categories. This reversal in the direction of the effect at the low end of the rating scale may indicate measurement error in the interviewers' assignment of the 0-9 category. As shown in figure 4.2, interviewers used the 0-9 rating more than the 10-19 or 20-29 ratings. It may be that some of the interviewers did not make full use of the range of the lower end of the scale (as evident from the gaps at the low end of the scale in figure 4.5) and rated cases as or near zero that were actually more likely to cooperate than the cases rated between 10 and 29.

Also contributing to the reversal in direction of the effect is that the cases rated the lowest are either not called again or, if called, no further contact is made (e.g. because the target person is screening his/her calls to avoid participating). The effect of this would be that most of the negative outcomes that would result on the next contact are never recorded in the data. As a consequence, the small number of cases in the 0-9 category that are contacted again would appear more highly correlated with cooperation than they actually are. An investigation of the last contact for each case shows that 70% of those contacts rated as 0-9 at the end of a contact were never contacted again. This is a much higher percentage of cases not contacted again than any other category (the next highest is the 10-19 category with 41% not contacted again, followed by 20-29 with 23% not contacted again). As shown earlier, in figure 4.3, most of the contacts rated between 0 and 9 were refusals, some of which would not have been attempted again. Those attempted again most likely screened their calls and avoided further contact. Of the small percentage that did have another contact, the success rate is higher (see figure 4.4). Therefore, relative to the 0-9 category, the 10-19 or 20-29 categorized likelihood ratings look less likely to cooperate.

Comparing the Classic model to the Classic+ model, the addition of the ratings does little to change the effect and significance of the covariates from the call records. There is some noticeable fluctuation in the coefficients for the number of appointments, with the ratings explaining some of the effect of the appointments. Also the effect of having reached the target person on the prior contact becomes marginally significant when the ratings are included in the model.

Interestingly, the significance and direction of the rating categories in the Classic+ model are opposite those in the Ratings-only model. Once all of the Classic model covariates are included, the significant or marginally significant likelihood rating categories are now below 70. This loss of the predictive power of the higher ratings indicates suppression -- the Classic model covariates share a lot of the same information with the ratings and better predict the cases most likely to cooperate on the next contact. The inclusion of the Classic model covariates also results in all categories of the likelihood rating have odds ratios less than one, compared to the lowest category (0-9). An examination of the strengths of association between the categorized ratings and covariates in the Classic model finds that the two variables characterizing prior refusal are strongly correlated with the ratings (refused on any

previous contact, Cramer's V=0.59; refused on the contact immediately prior to the current contact, Cramer's V= 0.74). It seems that these refusal indicators are explaining cooperation for the lowest likelihood category. With (lack of) cooperation in the 0-9 category explained by refusal history, the few non-refusing case appear to be very successful, making this category appear to be more predictive of cooperation than any other category. If these refusal indicators are dropped from the model (not shown), the odds ratios more closely resemble those in the Ratings-only model with the ratings of 50 and 70 and higher being more predictive of cooperation than the lowest category (although not significant). Although this test provides a plausible explanation for the unusual strength of the lowest likelihood category, unfortunately, the strong correlation between the ratings and refusal history makes interpretation of the coefficients of the likelihood ratings in the Classic+ model illogical.

Beyond a simple examination of the significance of the coefficients, a likelihood ratio test (without the random effect for interviewers) concluded that the Classic model is significantly improved when the two likelihood rating variables, the categorized rating and the average rating, are added ($\chi^2(12) = 67.1$; $p<0.0001$). In addition, including the average of the likelihood ratings significantly improves a model using only likelihood ratings and no other covariates ($\chi^2(1) = 5.78$; $p=0.016$).

Table 4.3 shows the fit (assessed using the pseudo R-squared and the AIC) and discrimination (assessed using area under the ROC curve) of the three propensity models tested, without the random effect for interviewers. Although the fit and discrimination of the Ratings-only model are not as good as the Classic model, the Classic+ model is significantly better than the Classic model which does not have the ratings. Hosmer and Lemeshow (2000) note that an ROC curve with an area of 0.70 or higher has "acceptable" discrimination (p. 162). Using this as a guide, areas that were between 0.50 and 0.70 were labeled as "minimally" discriminated.

Table 4.3. Fit and Discrimination for the Three Propensity Models (models run without random effects)

Model	n	Pseudo R-squared	AIC	AIC df	Area under ROC curve	Assessment of discrimination
Classic	3091	0.0875	3308	25	0.7053	Acceptable
Ratings-only	3091	0.0627	3394	24	0.6703	Minimal
Classic+	3091	0.1062	3265	37	0.7187	Acceptable

When the predicted propensities estimated from each of the models for each of the contacts are grouped into three tertiles, representing low, medium, and high propensity groups, a cross-tab of the Classic and Classic+ models shows that 78% of the contacts have the same categorized predicted probability in both models. The lowest tertile has the most agreement, with 87% of the Classic model contacts also classified as low propensity in the Classic+ model. The middle tertile has the lowest level of agreement, at 67%. The difference between the distributions indicates that the models are not equivalent and disagree on the classification for over 20% of the contacts. However, the comparison does not indicate if one model is consistently better at predicting cooperation than the other.

The net reclassification improvement (NRI) quantifies the movement between categories to calculate an overall improvement in the model when the ratings are included. Table 4.4 below shows the movement between propensity tertiles when comparing the Classic model to the Classic+ model. This is separately examined for contacts that result in cooperation (respondents) and those that do not (nonrespondents). The NRI equals 3.0%, indicating that the reclassifications made when the likelihood ratings are included in the model improve the accuracy of the predictions, even when broadly categorized into tertiles. The positive NRI is mostly attributed to more respondents that were reclassified into a higher propensity tertile than a lower one in the Classic+ model. Although the improvement is positive overall, some of the improvement is negated by the reclassification of cases into a tertile further from the truth.

Table 4.4. Percent of Contacts with a Corresponding Follow-up Contact in Each Propensity Tertile that are Reclassified when Comparing the Classifications from the Classic Model to the Classic+ Model, Shown Separately for Cases that Respond on the Current Contact and Those that Do Not

	Respondents	**Nonrespondents**
	N=817	N=2,274
Tertile change	(%)	(%)
Lower probability	10.0	11.4
No change	77.7	78.0
Higher probability	12.2	10.6
Overall improvement	2.2	0.8

4.4.3 Propensity modeling in a responsive design context

Following the assessment of the models using all contacts, the analysis investigates the performance of the models in a responsive design context, which involves analysis of the contacts up to a particular date of data collection. For each date, probabilities of cooperation at the next contact were estimated for the three propensity models at the close of twelve different dates of data collection. Using the predicted probabilities from the most recent contact for each case (up to the date of the "daily" model), cases were classified as high, medium or low propensity cases for each type of model (i.e., Classic, etc.). The boundaries for each tertile, for each date that the "daily" model was run, are shown in figures 4.6 - 4.8 for each model (the corresponding data for the figures can be found in appendix 4I). Generally, there is a very slight decrease in the distribution of propensities as data collection progresses and a widening of the range of high propensity cases.

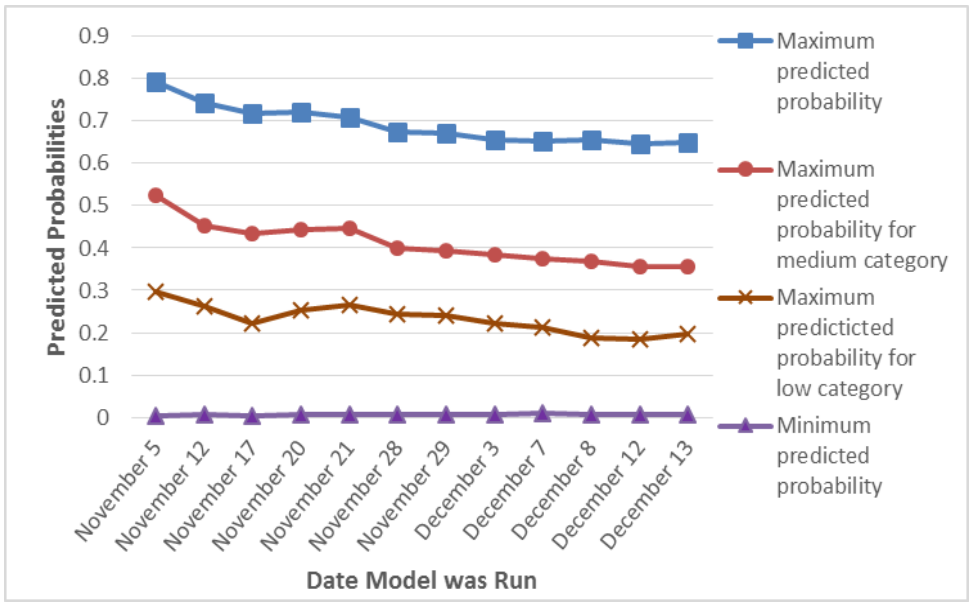


Figure 4.6. Boundaries of the Predicted Probabilities of Each Tertile Estimated from the Classic Model

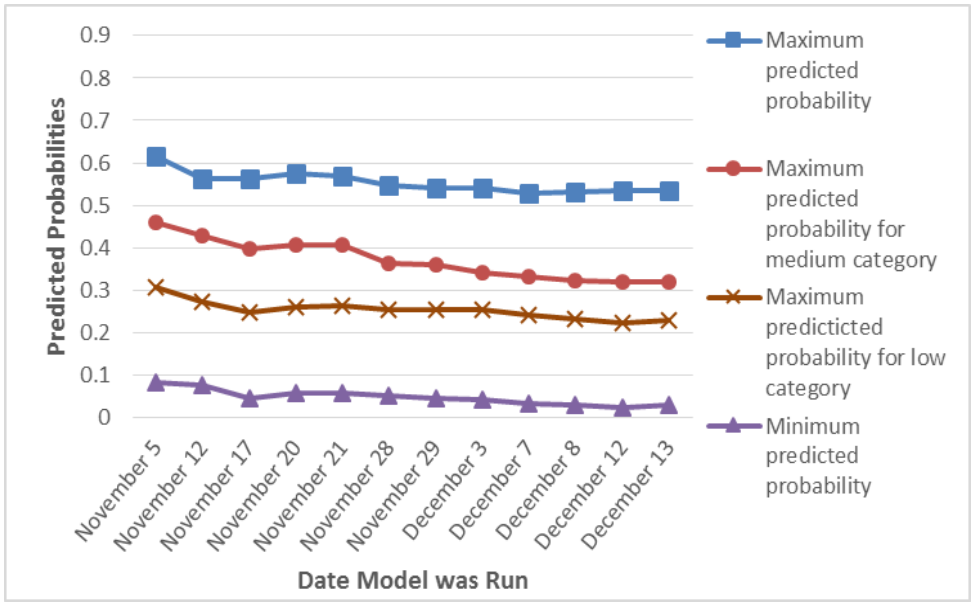


Figure 4.7. Boundaries of the Predicted Probabilities of Each Tertile Estimated from the Model with Likelihood Ratings Only

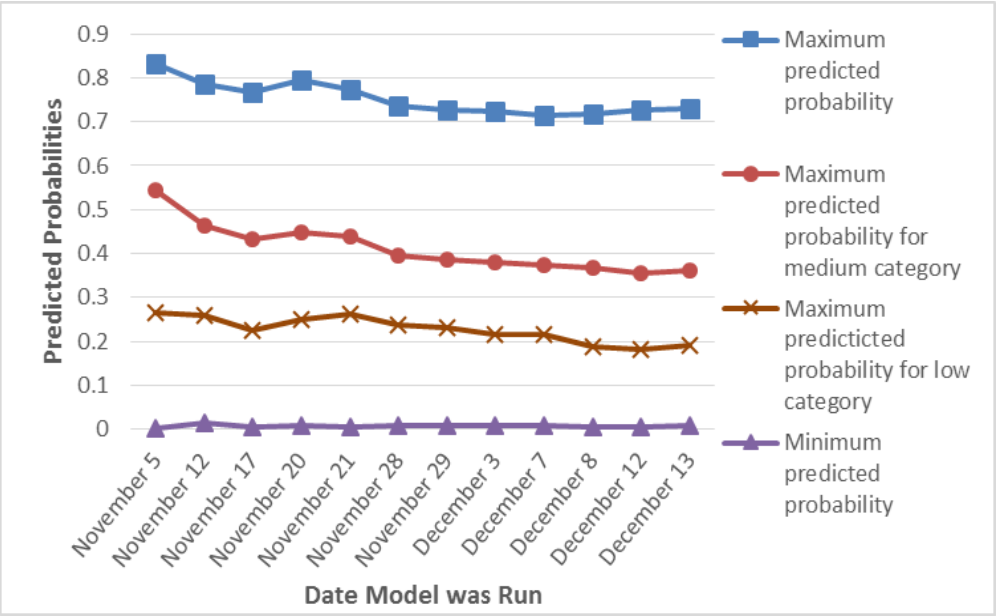


Figure 4.8. Boundaries of the Predicted Probabilities of Each Tertile Estimated from the Classic Model with Likelihood Ratings Added (Classic +)

For the analyses in the responsive design context, only active cases (those that have not yet cooperated) are examined because predicting cooperation for cases that have already cooperated is not useful for fieldwork management. The maximum number of probability estimates per case is twelve, corresponding to the number of days that were monitored. To illustrate the range of predicted probabilities an active case can have during the data collection period, figure 4.9 shows the propensities for the first 25 cases, estimated from the Classic model for each date monitored. Some cases, such as case 14, only have one probability estimate because they provided an interview at a subsequent contact on or before the next daily monitoring date after the second contact. Other cases, like case 48, have multiple estimates because they remained active for a larger portion of the field period than other cases (i.e. the case did not cooperate for a long period of time or at all after the second contact). Finally, there are cases not depicted in figure 4.9. This is because either the case cooperated on the first contact and is not in the analysis (e.g. 11 and 12) or the case cooperated on the second contact and is not an active case of the purposes of daily monitoring (e.g. cases 4 and 10). This figure also shows that the probability ranges across the data collection period for some cases can be narrow (e.g., case 25) or wide (e.g., case 5). To clearly describe when a case is part of the daily propensity modeling or not, appendix 4J contains detailed case histories for a selection of cases shown (or not shown) in figure 4.9.

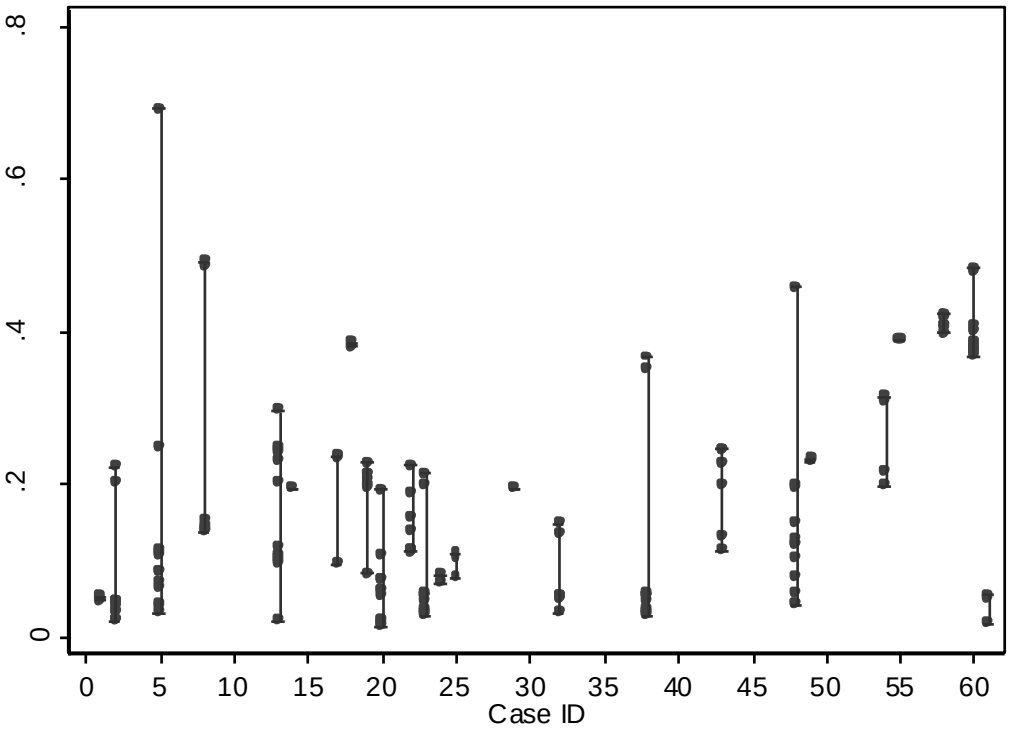


Figure 4.9. Predicted Probabilities Estimated from the Classic Model for the Dates Monitored for the First Twenty-Five Cases

To evaluate whether the likelihood ratings noticeably change the categorization of a case on a daily basis, the tertiles assigned to the probabilities generated from the Classic model were compared to the tertiles for the Classic+ model. Figure 4.10 shows the percent of “active” cases each day that have the same categorization in both models. There is little fluctuation in the agreement, with agreement between 83% and 89% for each date monitored. This seems to indicate that the Classic and Classic+ models are generally classifying active cases into the same category (high, medium or low probability), even early in the field period.

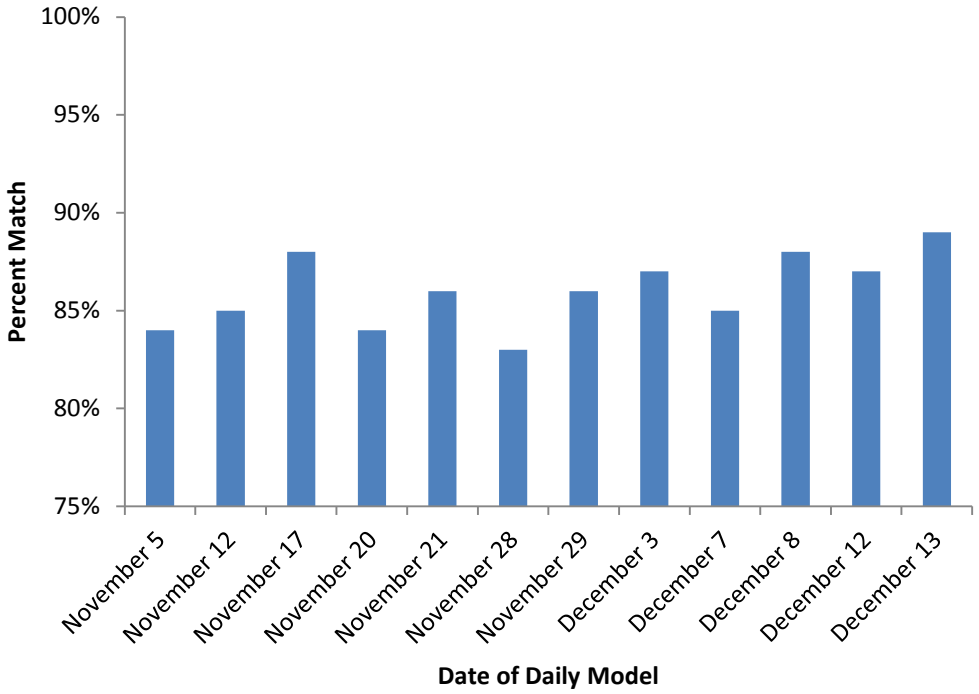


Figure 4.10. Percent of Cases Assigned the Same Category for Both the Classic Model and the Classic Model with the Likelihood Ratings (Classic+)

To examine the ability of each model to predict cooperation or non-cooperation at the next contact for the dates selected, the percent of active cases that cooperated at the next contact was calculated for each probability tertile. These percentages represent success rates for these tertiles on the dates of monitoring and can be compared across the three models. Ideally, the high probability tertile should have the highest percent of cases that cooperate on the next contact, the low tertile should have the lowest percentage, and the medium should be somewhere in between. A summary of these percentages across all dates monitored is presented in table 4.5. When aggregated this way, the expected trend appears most clearly for the Classic+ model and Ratings-only models. Although the high probability tertile in the these models has the highest percent of cases that cooperate on the next contact, the success rate itself (10% and 12%) is not high. Comparing the performance of the three models within a tertile, there appears to be little difference between the success rates for the low probability tertiles (all models predict that 5% of the cases cooperate). However, the Ratings-only model appears to predict success better than the Classic and Classic+ models for the high probability tertile (12% compared to 7% and 10%).

Table 4.5. Success Rate of High, Medium, and Low Probability Tertiles to Predict Cooperation on the Next Contact

	High probability cases		Medium probability cases		Low probability cases	
Model	Number of active cases in tertile for all dates monitored	Percent cooperated on next call	Number of active cases in tertile for all dates monitored	Percent cooperated on next call	Number of active cases in tertile for all dates monitored	Percent cooperated on next call
Classic	471	7%	1122	10%	2776	5%
Ratings-only	554	12%	1305	6%	2510	5%
Classic+	410	10%	1218	9%	2741	5%

Although the aggregate success rate is informative, the true test of these models is in the context of a live responsive design by evaluating the accuracy of the predictions for each date separately. Figures 4.11, 4.12, and 4.13 show the daily success rates for the high, medium, and low probability groups for the dates monitored. Detailed data corresponding to the success rates shown in the figures can be found in appendix 4K. The success rates for the low probability tertile are generally the same across the three models (figure 4.13). It is difficult to interpret which model performs better in the medium probability tertile because there isn't an expectation that the success rate will be high or low. Nonetheless, figure 4.12 shows that the Ratings-only model has a consistently lower success rate than the two other models, revealing that this model classifies medium tertile cases as less likely to respond than the other models. The success rates are also different for the high probability tertile (figure 4.11). The model using the likelihood ratings only appears to be more successful at predicting cases that will cooperate on the next contact compared to the other two models. The difference is particularly noticeable towards the end of the data collection period, when identifying the most likely to cooperate cases is more difficult. Therefore, the likelihood ratings appear to be valuable for predicting the cases most likely to cooperate, especially towards the end of the field period, and their strength may be hampered by the other covariates in the classic model, as demonstrated by the lower success rate of the Classic+ model.

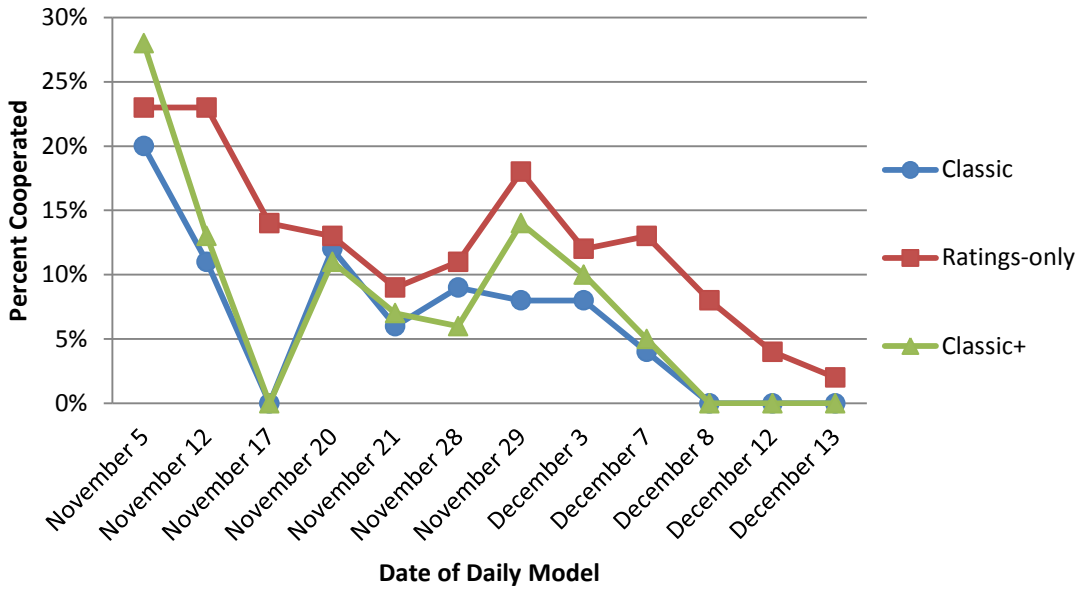


Figure 4.11. Percent of High Probability Cases that Cooperated on the Next Contact for Each Model for Each Date Monitored

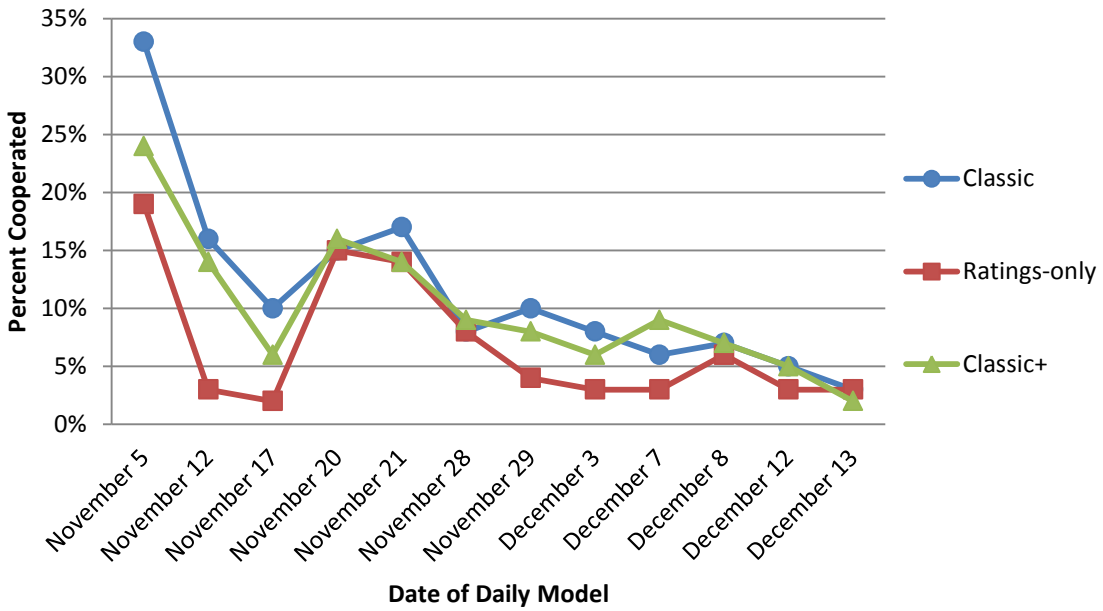


Figure 4.12. Percent of Medium Probability Cases that Cooperated on the Next Contact for Each Model for Each Date Monitored

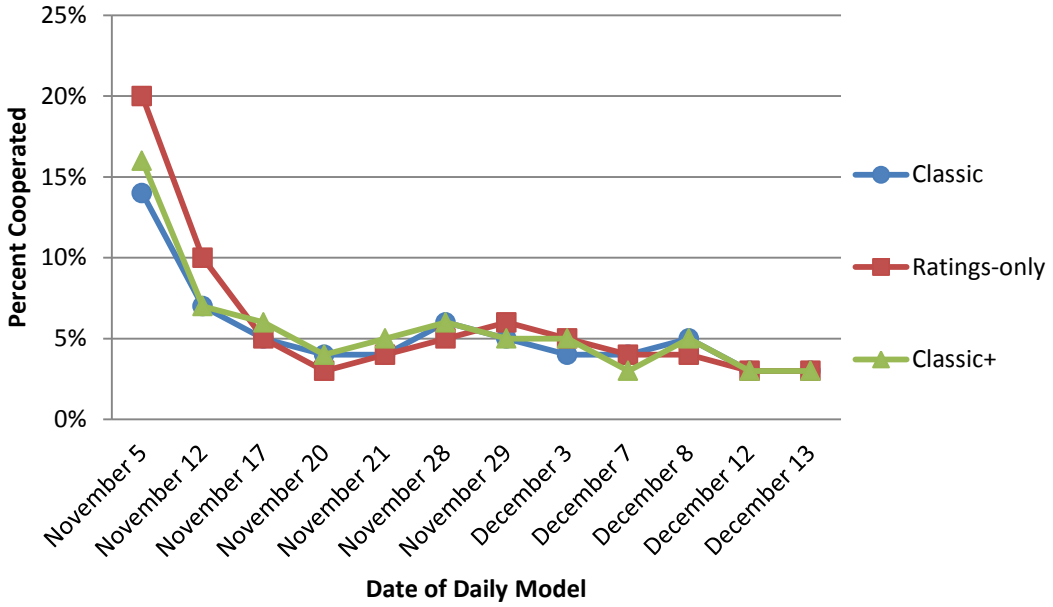


Figure 4.13. Percent of Low Probability Cases that Cooperated on the Next Contact for Each Model for Each Date Monitored

4.5 Discussion

This analysis uses discrete time logistic hazard models, controlling for the random effect of interviewers, to evaluate whether an interviewer observed rating of the respondent's likelihood to respond, recorded at each contact, could improve a "Classic" response propensity model including call record and interviewer characteristics. The results show that the ratings are significant in the Classic model (creating the Classic+ model), and the fit and discrimination of the Classic model was notably improved when the ratings were added. When the models were used in a responsive survey design context, which involves daily monitoring during data collection, a model with the likelihood ratings-only appears to outperform both the Classic and Classic+ models when predicting cooperation of the high probability cases, especially during the end of the data collection period. Although the ratings on their own have weaker fit and discrimination statistics overall when compared to the Classic model, it appears that for the specific application tested (responsive survey design) and the subgroup of interest in this application (high probability cases), the interviewer recorded likelihood ratings may be more useful than the Classic propensity model with only call record data.

Although the success of the ratings in this analysis is laudable, there are some flaws in the design of the rating that, once addressed, could further improve their performance. The interviewers were asked to make a rating about the future which is difficult in general but is made more challenging by the fact that the same interviewer will likely not make the next contact. Contributing to this is that the question does not specify whether the interviewer should answer the observation based on his ability or his perception of other interviewers' abilities to secure cooperation. The current interviewer observation question presupposes that the interviewers on this study have the same level of ability in terms of securing cooperation, and securing an interview only depends on the differences between the cases. The models show a small but significant interviewer effect that cannot be explained by interviewer characteristics.

This statistic confirms that not all interviewers have the same level of ability and the predictive power of the ratings is likely to suffer because of this.

As demonstrated by the descriptive statistics and the recategorization of the scale, the response options for this question may not be ideal. The distribution reveals significant rounding and heaping, indicating that the interviewers use the scale in a categorical way. The scale would probably be better as 10 points or less. Additional research can be done to fine tune the number of categories and decisions regarding the middle category. Another issue with the scale is that it asks for a probability, which is already difficult for most people to understand (Tversky and Kahneman 1974), but the task is potentially further complicated by asking for the probability in terms of a percentage. However, the data collection agency felt that their interviewers would understand the scale in terms of percent and therefore, prior experience of the interviewers can be a factor in the scale and question design.

When the response scale was categorized and then introduced into the model as dummy variables, information was lost in two ways. First, the categorization clustered the responses and second, the dummy variables did not preserve the order of the response options. Although one solution would be to keep the interviewer rating as continuous, this was not favored, given the distribution. Another solution would be to apply current nonparametric methods of modeling in the form of generalized additive models (Wood 2006) which are touted to provide a better fit to the data. One technique, demonstrated in Tutz and Gertheiss (2013), preserves the relative relationships of the ordinal responses by introducing a penalty that “fuses” or collapses categories that are essentially the same with respect to cooperation, thereby clustering the response options according to the fit of the data. The modeling method used was chosen because it is commonly used in responsive survey design, the application that this analysis aims to improve. However, future developments of responsive survey design techniques could explore the implementation of generalized additive models.

Due to difficulties finding a reasonable imputed value of the rating for the first contact, the first contact was dropped from the analysis. This resulted in not only a loss of data (and power) but the analysis is examining a different group of people: in addition to the sampled persons who were never contacted, those contacted just once are not part of the analysis. Removing those who cooperated on the first contact is not a significant concern in this context since a researcher running daily propensity models to direct fieldwork is not interested in these cases. However, sampled persons who were contacted once, did not provide an interview, but were never contacted again are of interest. Either these people are passive refusers who are avoiding contact or the interviewer or agency is doing a poor job of gaining contact. Both scenarios are of interest to a fieldwork manager.

Although the ratings significantly improve the Classic model, their power is diminished by the correlations with other covariates, notably, the refusal outcome. Not only do the ratings and call outcome characterize the case at the same point in time but also the interviewer determines the codes for both data. One could argue that using both the call record data in the Classic model and the likelihood ratings are not necessary and only one type of data should be applied in the models. I am cautious to recommend the use of likelihood ratings over the call record data, even though the ratings appear to be more predictive of cooperation among the high probability cases, because none

of the discrimination statistics or success rates of the models are good enough to strongly recommend any one of them. As noted in the introduction, there is a need to improve the performance of response propensity models, especially as survey researchers become more dependent on the predictions from these models. Either finding or creating new forms of paradata or improving the quality of existing paradata is necessary to achieve this. Clearly, this analysis of a new form of paradata contributes to that effort but additional work is necessary for interviewer recorded likelihood ratings to make a stronger impact on the model performance.

A final note on the findings is that although the dataset allows for an interpenetrated design of interviewers, the results of the analysis are less applicable for CATI than CAPI[22] because an autodialer is going to send the cases to the field even if they are low priority. Making additional calls is inexpensive and interviewers already scheduled to work need to be kept busy. Unless the caseload in the call center is very high, the cost savings of predicting which cases are likely to be cooperative at the next contact is more useful in a CAPI data collection. Therefore, the methods used here should be duplicated to assess the usefulness of these ratings in that setting.

[22] Computer Assisted Personal Interviewing

Chapter 5: Conclusion

5.1 Summary of work presented

In the analyses presented, I address three main questions that build on one another, each furthering the exploration of the accuracy and utility of interviewer observations for nonresponse applications such as weighting and responsive survey design. First, I assess the magnitude of the measurement error in several commonly collected interviewer observations and identify correlates of that error. Next, I investigate how the magnitude of error in the observations compares to the magnitude of error in another type of data, commercial data, that is used for the same nonresponse application. Lastly, I evaluate the impact of the measurement error in the observations on a specific nonresponse application, responsive survey design. Together, the three papers provide a coherent body of work on the quality and utility of interviewer observations for nonresponse applications.

The results show that for the five observations analyzed using the UK Census data, the measurement error is minimal, with accuracy ranging from 87 to 98 percent. The conclusions postulate likely reasons as to why the accuracy is not higher for some of the observations, providing workable solutions for reducing the measurement error. Correlates of accuracy found in the data pertain to the visibility of the property, such as whether the housing unit is a standalone home or an apartment in a multi-unit structure, and the level of interviewer-respondent interaction, indicated by the result code. These correlates as well as the significant interviewer effects represent sensible mechanisms of the influences on measurement error.

When interviewer observations designed to match key survey outcomes are compared to typically available commercial data in the German PASS study to determine which is the better predictor of key survey outcomes, the analysis finds that interviewer observations are more predictive, particularly for the special subpopulation that this survey targets. This result favors the use of interviewer observations for nonresponse weighting over another form of commonly used data, even with measurement error. Combining this finding with the conclusions from the first paper, adjustments to interviewer training and protocol designed to improve the observations may further reinforce the worth of observations for nonresponse applications.

Looking to improve nonresponse applications that require accurate prediction of a case's propensity to respond, a new kind of observation, taken at the call level, was designed to capture the likelihood of a case to respond. Correlational analyses revealed that the interviewers' ratings of likelihood are predictive of cooperation, despite the subjectiveness of the observation. Multivariate analyses find that the performance of the response propensity models is significantly improved when the likelihood rating is included in the model, especially at the end of the field period. This finding supports the creation of new observations for specific nonresponse applications such as responsive survey design. However, as in the measurement error analysis of the UK observations, this new observation would benefit from improvements, especially pertaining to the design of the question and response options.

Across all three analyses, the results are encouraging. The interviewer observations have shown to be useful for nonresponse applications such as weighting and responsive survey design and, where measurement error is notable, workable solutions are available. Therefore, investing in the improvement and development of interviewer observations holds promise. Key findings from the work presented caution survey researchers to design observations that first, interviewers are capable of making (e.g. observing the presence of young children rather than children up to the age of 18) and second, accurately capture the construct(s) of interest. This second lesson was evident in the lower accuracy of the observation of council housing in the UK data as well as the better performance of the PASS interviewer observations over the microm data for the prediction of key survey outcomes. Related to this point, the findings also advise researchers to design interviewer observations with the same care applied to survey questions and response options. In particular, the performance of the likelihood rating could be improved by better question design.

5.2 Future research

Given the promising findings of these analyses and the recommendations for improvement of the interviewer observations, the next steps in this line of research are to make deliberate attempts to improve the quality of the observations and then reevaluate their performance in nonresponse applications using methods similar to those presented here. The expected result is a reduction in measurement error and improved prediction of key survey outcomes and response propensity.

Following or in parallel to these efforts, there are other research agenda items that could be pursued. One research area not emphasized in this work is the reduction of interviewer effects in the observations. To tackle this problem, researchers could explore and document the cognitive process interviewers undergo when making interviewer observations. The cognitive process for interviewers could generally follow the cognitive process that has already been developed for survey respondents which includes: comprehension of the question, retrieval of relevant information, judgment and integration of the information, mapping onto a response, and possibly editing that response (Tourangeau et al. 2000). By thoroughly studying each of these steps in the context of making observations, researchers could document the cognitive difficulties when “answering” observation questions. Understanding these difficulties would provide essential insight into how measurement error arises in interviewer observations.

A second outcome of studying the cognitive process could be the construction of a unique cognitive process tailored to account for the differences between making observations compared to answering a survey question. For example, survey respondents are susceptible to social desirability when providing a response to a sensitive question. From my experience, interviewers are subject to a social pressure that is tangential to this but not quite the same that inhibits them from recording an unflattering judgment on someone that they do not know. Therefore, when interviewers are asked to record whether someone is in a sexually active relationship or not (as is done on NSFG), some may be hesitant to record that someone is sexually undesirable, even if that is their judgment from what is observable. On the other hand, some interviewers have reported sensitivity to recording an observation at all (regardless of the answer marked on the observation form), stating that they feel self-conscious admitting that they judged someone on something so personal. Nuances of the process such as these help identify what observations interviewers are capable of making,

which is one of the key cautions mentioned above. Even if visible evidence is available to make an observation, interviewers introduce a human element to the process. This human element may prevent the reporting of a judgment or compel the editing of a response. If problematic observations are essential, researchers must investigate solutions to desensitize interviewers and prevent the triggers in the cognition process that result in measurement error.

Related to this recommendation, features such as training and question design should be studied to understand how they impact the stages of the cognitive process when making interviewer observations. Taking just the comprehension stage, experiments with question wording and response options would reveal how well a question should be designed in order to minimize comprehension error. Interviewers are more accustomed to hearing and answering survey questions than survey respondents and may not need as precise question development. Interacting with this is the effect of training – it may be that sufficient training on the observations overcomes most or all design flaws in the observation questions. Another element that is unique to the collection of interviewer observations is that these questions are "asked" of the interviewer repeatedly, as they visit more addresses. As the field period progresses, interviewers may (possibly quickly) reach a point when they do not read the observation questions anymore, eliminating the need for meticulous question design. Also, as interviewers make more observations, their comprehension of the question may change, given the experience and information that they have gathered from prior addresses. Survey researchers may need to prevent this change in comprehension by providing all possible scenarios when training.

Besides the need to understand the interviewers' cognitive process when recording observations and the elements that may interact with that process, an additional gap in understanding the quality of interviewer observations is the documentation of *how* interviewers collect observations. The research thus far on interviewer observations has not detailed the routines and procedures interviewers execute when in the field. Although interviewers are trained to follow certain protocols when recording observations, they may not do so, or not at least consistently. If they do follow the protocol, there are still behaviors that are not specified in the instructions but could affect the measurement error. For example, most surveys instruct interviewers to collect the area observation on the first call to a household before contact is made. Interviewers can follow this instruction in several ways: making a special visit to the property before attempting to make contact, sitting in their cars outside the house to specifically record the observations before attempting contact, simply mentally noting the observations before they make contact but recording them after the visit, etc. All of these routines could have different implications for measurement error and researchers should understand this and recommend particular procedures. The investigation of this mechanism is next on my research agenda and I have applied for a small grant to study this.

The ideas outlined above recognize that there are still some gaps in understanding the measurement error and overall quality of interviewer observations. However, the results from the analyses presented provide a substantial foundation of knowledge on which to build further research. Interviewer observations have proven and will continue to prove their worth for nonresponse applications; investing effort to improve the quality of interviewer

observations will allow researchers to benefit from the full potential of this useful form of paradata.

Appendices

Appendices for Chapter 2

Appendix 2A: Further Information on Missing Data in the Interviewer Observations

For each observation, an interviewer response is recorded in the data. However, interviewers had a “Don’t Know” option (in the Council observation, this is labeled “Unable to code/NA”) for all five of the analyzed observations. Since the “Don’t Know” responses do not allow for the assessment of accuracy, they either need to be imputed, treated as incorrect, or removed from the analysis. Each option has drawbacks and I felt that the least amount of error would be introduced by dropping these cases. I recognize the possibility that, by eliminating these responses, the accuracy rates could be overestimated, assuming the level of error in these would-be observations is higher than in those analyzed. Because the missing data rates are very low for most observations, this effect is likely to be minimal.

Relevant to any discussion on missing data is its prevention. This is especially important if the observations are to be used for correction of nonresponse bias. If an interviewer observation questionnaire is well-designed, providing missing response options to the observation questions can be avoided. For example, the Census observation form did not need to include the “Don’t Know” option, as the skip instructions ensured that all observations were relevant. In addition, the use of automation (CAPI, as opposed to PAPI) should prevent unacceptable missing information. However, as mentioned above, forcing interviewers to make a guess may lead to lower quality observations. Further analyses comparing the accuracy of interviewer observations that interviewers are confident about versus not confident about would inform us as to the value of forcing responses and the resulting effect on nonresponse adjustment.

Table A1. Missing Data Rates Overall and within Result Code Category for each Interviewer Observation

	Overall		Cooperative (n=13,446)	Refusals (n=3,608)	Noncontacts (n=817)
Observation	**n**	**Percent Missing**	**Percent Missing**	**Percent Missing**	**Percent Missing**
Type of HU	17,871	0.3%	0.2%	0.5%	1.2%
Council	17,871	4.4%	3.5%	6.7%	9.2%
Working	17,054	8.6%	1.2%	36.2%	--
White	17,054	1.5%	0.3%	6.0%	--
Children	17,054	12.0%	6.9%	31.3%	--

Note: The percentages presented are specific to the result code and the observation. For example, 1.2% of the 13,446 cooperative cases were missing for the Working observation. Among these same cases, 0.3% had missing data for the White observation.

Appendix 2B: Comparison of the True Values from Census Self-Reports to the Interviewer Observations

Table B1. Cross Tabulations Showing the Prevalence of each Characteristic in the Analysis and the Percent Agreement between each Interviewer Observation Category and the Census Data

	Interviewer Observation		
Census Report	**House**	**Not a House**	**Total**
House	81.9%	1.3%	83.1%
Not a House	1.6%	15.2%	16.9%
Total	83.5%	16.5%	(n=17,815)

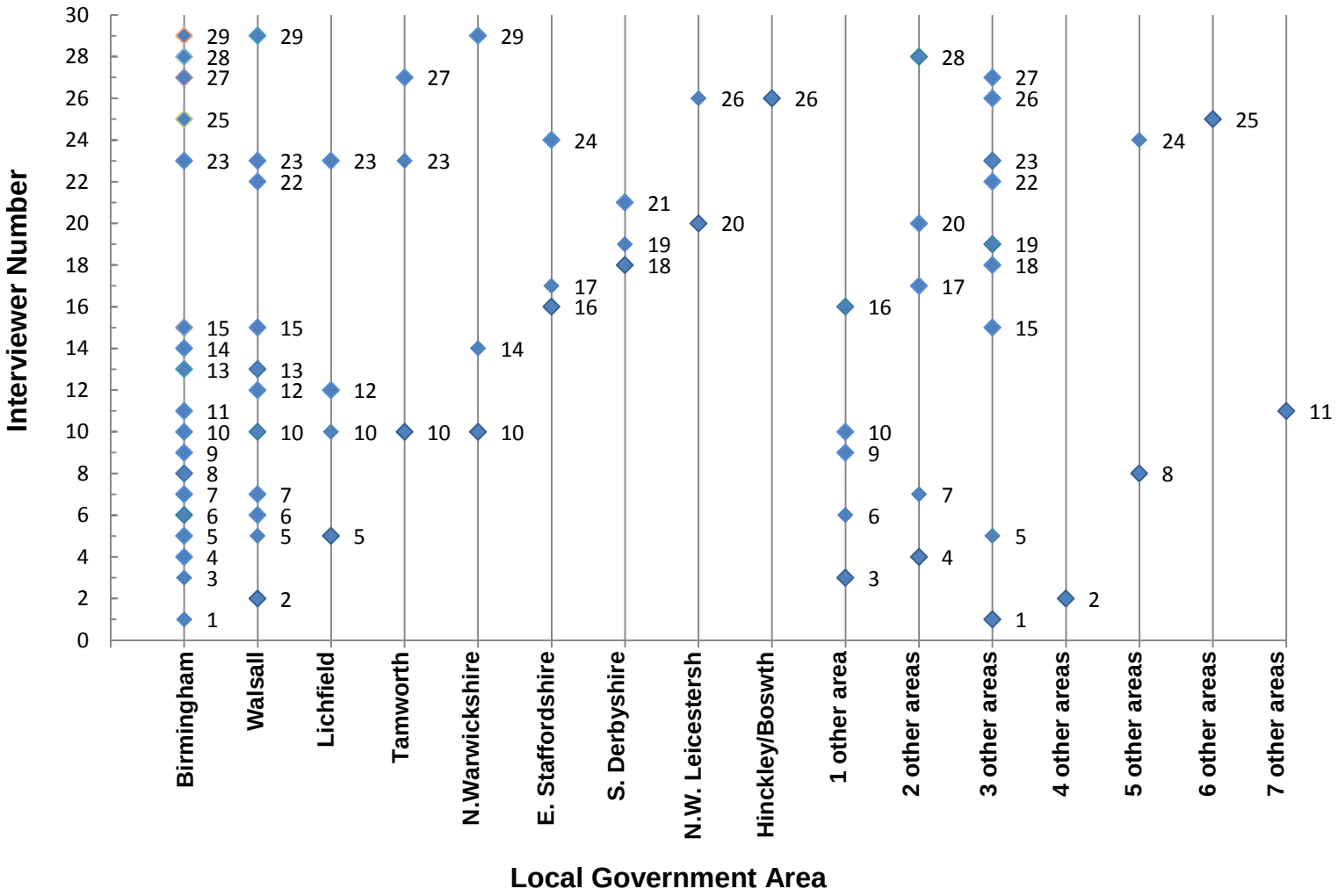
	Interviewer Observation		
Census Report	**Council House**	**Not Council House**	**Total**
Council House	14.9%	4.6%	19.5%
Not Council House	8.7%	71.8%	80.5%
Total	23.6%	76.4%	(n=17,069)

	Interviewer Observation		
Census Report	**Working Adult**	**No Working Adult**	**Total**
Working Adult	60.4%	3.2%	63.6%
No Working Adult	4.0%	32.4%	36.4%
Total	64.4%	35.6%	(n=15,575)

	Interviewer Observation		
Census Report	**All White**	**Not All White**	**Total**
All White	93.2%	0.8%	94.0%
Not All White	1.3%	4.7%	6.0%
Total	94.5%	5.5%	(n=16,784)

	Interviewer Observation		
Census Report	**Children**	**No Children**	**Total**
Children	26.4%	5.1%	31.5%
No Children	2.0%	66.5%	68.5%
Total	28.4%	71.6%	(n=14,965)

Appendix 2C: Illustration of the Cross-Classification of the Data



Note: Since interviewers working in the nine chosen areas also worked in areas outside of those nine, "other area" categories were added, designating how many other areas the interviewers worked. One can then determine how many areas each interviewer worked in total.

Figure C1: Illustration of the Cross-Classification of Interviewers and Areas for an Extract of Nine Areas which Border each other, from the Case Base for the Type of Housing Unit Observation

Appendix 2D: Supplemental Information Concerning the Analysis of the Multilevel Models

Table D1 shows the effect of interviewers and areas on the accuracy of each of the five interviewer observation variables, using a multilevel analysis and adding groups of similar variables at each step, as described in the Methods section. In the empty two-level models only accounting for the effect of interviewers (shown in the top section of Table D1), the interviewer random effect variances are significantly different from zero for all observations. In the empty models accounting for only area effects, the variance attributed to area is significant for all observations except Children. For the most part, these significant results carry through to the empty cross-classified models accounting for both interviewer and area effects simultaneously, indicating that interviewers and areas do contribute to the variance in the accuracy of the observations. The exceptions are the models predicting accuracy of the observation of Working, where the cross-classification removes the effects of both interviewers and areas, and White, where the interviewer part of the random effect is no longer significant.

The DIC, an indicator of model complexity and fit (see Spiegelhalter et al. 2002), is reduced for most models when the cross-classified model is introduced indicating that controlling for both interviewer and area effects simultaneously improves the fit of the models. The difference between the two-level model accounting for random interviewer effects and the cross-classified model with both interviewer and area effects is largest for Type of HU. The Council model shows the most significant reduction in DIC between the two-level model for area and the cross-classified model. Working shows virtually no reduction between the empty two-level model and the cross-classified model.

As covariates are added to the cross-classified models, the random effects of interviewers and areas on the accuracy of the observations are gradually explained by household, interviewer and area characteristics. The first step of the cross-classified model development, inclusion of the true value of the observation from the Census data, reduces the effect of both interviewers and areas across almost all models, and eliminates the significance of area in the Type of HU and White models (area was already not significant in the empty Children and Working models). Therefore, for all models except Council, the area effect is no longer a concern beyond the introduction of the true value indicating that, in addition to affecting the measurement error, the true value varies by area. It is also noteworthy that a large drop in the DIC occurs across all models when these variables are included.

Unlike the area effect, the interviewer effect is more gradually explained as the result code, survey indicator, significant household characteristics (such as ownership, type of structure, number of adults, and ethnicity), and interviewer characteristics (such as age, experience and attitudes from the interviewer survey) are introduced. Adding interviewer characteristics to the model fully explain the effect for both the Type of HU and Children observations (these effects were already not significant in the empty cross-classified Working model). Although the significance of the interviewer random effect fluctuates across the steps of the model development for the White observation, after all covariates are included in the model, the contribution of interviewers to the variance is still significant, but barely. For the Council

observation, the introduction of household, interviewer and area characteristics reduces the random effect slightly, but it remains significant throughout the stages of model development.

Once all covariates are entered, the interviewer effect is fully explained for three of the five models (Type of HU, Working, and Children) and the area effect for all observations except Council. The result is a significant influence of interviewers, after including household, interviewer and area characteristics, for two of the five observations, Council and White, and of areas for only the Council model. As neither effect is fully explained in the final Council model, the accuracy of this observation should be modeled using a cross-classified model. The other observations only require a two-level model, and Working does not necessarily require any multilevel analysis.

Table D1. Estimates of the Interviewer and Area Random Effect Variances with Standard Errors (in parentheses) and 95% Confidence Interval*, from the Stepwise Modeling Procedure Predicting Accuracy of the Observation, using MCMC in MlwiN†

	Type of Housing Unit			Council			Working Adult			White			Children		
	Interviewer Variance var (se)	**Area Variance** var (se)	**DIC**	**Interviewer Variance** var (se)	**Area Variance** var (se)	**DIC**	**Interviewer Variance** var (se)	**Area Variance** var (se)	**DIC**	**Interviewer Variance** var (se)	**Area Variance** var (se)	**DIC**	**Interviewer Variance** var (se)	**Area Variance** var (se)	**DIC**
Empty Models															
Interviewer Random Effect	**0.735** *(.139)* [0.482, 1.035]		4657	**0.696** *(.067)* [0.571, 0.835]		12756	**0.178** *(.058)* [0.060, 0.296]		8062	**1.022** *(.204)* [0.665, 1.455]		3349	**0.134** *(.051)* [0.046, 0.238]		7600
Area Random Effect		**0.508** *(.104)* [0.328, 0.735]	4646		**0.431** *(.056)* [0.345, 0.552]	12895		**0.068** *(.030)* [0.011, 0.133]	8077		**0.630** *(.139)* [0.389, 0.933]	3363		0.028 *(.028)* [0.000, 0.097]	7614
Cross-classified	**0.189** *(.096)* [0.013, 0.390]	**0.424** *(.105)* [0.242, 0.652]	4633	**0.271** *(.047)* [0.185, 0.372]	**0.330** *(.053)* [0.236, 0.443]	12753	0.037 *(.031)* [0.008, 0.139]	0.057 *(.037)* [0.001, 0.109]	8076	0.253 *(.153)* [0.003, 0.545]	**0.553** *(.142)* [0.300, 0.860]	3344	**0.102** *(.036)* [0.034, 0.178]	0.017 *(.020)* [0.001, 0.067]	7591
Cross-classified Covariates															
True value (TV)	**0.185** *(.090)* [0.026, 0.369]	0.118 *(.074)* [0.002, 0.301]	4265	**0.222** *(.042)* [0.147, 0.311]	**0.250** *(.047)* [0.165, 0.347]	12604	0.022 *(.027)* [0.001, 0.092]	0.051 *(.026)* [0.008, 0.109]	7892	0.194 *(.103)* [0.008, 0.419]	0.094 *(.067)* [0.006, 0.242]	2646	**0.173** *(.048)* [0.088, 0.276]	0.019 *(.020)* [0.001, 0.072]	6168
TV + Result Code (RC)	**0.176** *(.090)* [0.016, 0.367]	**0.116** *(.068)* [0.015, 0.267]	4256	**0.222** *(.043)* [0.145, 0.312]	**0.250** *(.045)* [0.169, 0.347]	12601	0.034*(.029)* [0.001, 0.104]	0.053 *(.028)* [0.008, 0.116]	7857	**0.205** *(.109)* [0.021, 0.442]	0.106 *(.076)* [0.002, 0.279]	2636	**0.162** *(.049)* [0.073, 0.265]	0.027 *(.025)* [0.002, 0.094]	6124
TV + RC + Survey (S)	0.103 *(.096)* [0.002, 0.309]	**0.136** *(.068)* [0.024, 0.285]	4253	**0.226** *(.043)* [0.147, 0.317]	**0.249** *(.045)* [0.171, 0.344]	12599	0.032*(.033)* [0.001, 0.113]	0.048 *(.032)* [0.003, 0.119]	7859	**0.220** *(.099)* [0.055, 0.435]	0.107 *(.084)* [0.001, 0.294]	2636	**0.150** *(.047)* [0.067, 0.249]	0.024 *(.023)* [0.001, 0.080]	6117
TV + RC + S + Household char. (HH)	**0.171** *(.083)* [0.029, 0.342]	**0.130** *(.069)* [0.019, 0.282]	4199	**0.208** *(.042)* [0.130, 0.296]	**0.239** *(.045)* [0.160, 0.332]	12337	0.012 *(.013)* [0.001, 0.052]	0.015 *(.016)* [0.001, 0.057]	7561	0.174 *(.125)* [0.001, 0.443]	0.112 *(.074)* [0.009, 0.284]	2612	**0.157** *(.047)* [0.070, 0.254]	0.027 *(.027)* [0.002, 0.098]	6011
TV + RC + S + HH + Interviewer char. (I)	0.093 *(.083)* [0.000 0.277]	0.115 *(.073)* [0.005, 0.281]	4177	**0.199** *(.042)* [0.124, 0.289]	**0.245** *(.045)* [0.162, 0.340]	12318	0.022 *(.022)* [0.001, 0.077]	0.009 *(.011)* [0.001, 0.040]	7555	**0.178** *(.108)* [0.014, 0.415]	0.100 *(.075)* [0.005, 0.287]	2596	0.079 *(.057)* [0.001, 0.197]	0.032 *(.028)* [0.002, 0.095]	5976
TV + RC + S + HH + I + Area characteristics	0.092 *(.072)* [0.004 0.258]	0.068 *(.050)* [0.003, 0.187]	4168	**0.188***(.041)* [0.112, 0.272]	**0.205** *(.041)* [0.133, 0.295]	12310	Not applicable			Not applicable			Not applicable		

*The confidence interval from 2.5% to 97.5% is shown in brackets under each variance estimation. If the lower bound of the confidence interval was less than 0.010, the bound was assumed to be zero and therefore, the variance was not significant. Using this criterion, significant values are bolded.

†The values in each cell are the point estimate (the means of 45,000 MCMC samples, with burn-in of 500 for all models except White which drew 120,000 MCMC samples). Standard errors (*se*) are calculated as the standard deviations of the estimates from the MCMC samples. DIC=deviance information criterion, an evaluation of model fit (see Spiegelhalter et al. 2002).

Appendix 2E: Wording of the Questions for the Interviewer Survey

Table E1. Wording of the Questions from the Interviewer Questionnaire which were Tested in the Multilevel Models

Question
How many years have you worked as an interviewer for SSD?
How many months have you worked as an interviewer for SSD?
What is your current SSD pay grade? *Interviewer; Advanced Interviewer; Merit 1; Merit 2; Merit 3; Field Manager.*
Have you ever worked for any survey organisations other than SSD?
For how many years did you work or have you worked as an interviewer for survey organisations other than SSD?
For how many weeks did you work or have you worked as an interviewer for survey organisations other than SSD?
Besides interviewing for SSD or other survey organisations, do you have any other paid employment?
The following questions use the response categories: Always, Frequently, Sometimes, Rarely, Never.
Before approaching the household, how often do you try to guess the type of people who are living in it?
Before a respondent has agreed to take part in a survey, how often do you ask to go into the home?
If you have just experienced a refusal, how often would you say it negatively affects how you feel about contacting the next household in your assignment?
If you have just experienced a refusal, how often would you say it negatively affects how you behave at the next household in your assignment?
The following questions use the response categories: Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree.
During the initial contact, it is more important to keep the conversation going than to seek a quick decision on participation from the household.
An interviewer should respect the privacy of the respondent.
I can easily use a wide variety of doorstep approaches.
I use a set structure for my doorstep approach.
I find it difficult to modify my doorstep approach even if I feel the situation calls for it.
Are you happy to carry out interviews on every weekday in accordance with your contracted number of days?
Are you happy to work in the evenings regularly?
What is your date of birth?
What is the highest educational qualification you have obtained? *Higher degree and postgraduate; Degree or degree equivalent; Other higher education; A levels or vocational level 3; O levels or GCSE grade A-C; Qualifications below the above; Trade apprenticeship/secretarial; Other qualifications- level unknown; No qualifications*
Are you: *Male Female*.

Appendix 2F: Estimated Coefficients for the Final Two-Level and Cross-Classified Models Predicting the Accuracy of Each Observation

Table F1. Estimated Coefficients and Significance for the Final Two-Level Models, with Random Interviewer Effects, showing all Covariates used to Predict the Accuracy of each Observation (estimated using Stata)

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
True Values										
House	--	--	--	--	--	--	--	--	--	--
Flat	-1.89	0.000	--	--	--	--	--	--	--	--
Caravan, Other	-2.62	0.000	--	--	--	--	--	--	--	--
Council House	--	--	-0.71	0.000	--	--	--	--	--	--
Working Adult	--	--	--	--	0.86	0.000	--	--	--	--
All White	--	--	--	--	--	--	3.36	0.000	--	--
No Children	--	--	--	--	--	--	--	--	--	--
1 Child	--	--	--	--	--	--	--	--	-2.91	0.000
2 Children	--	--	--	--	--	--	--	--	-1.38	0.000
3+ Children	--	--	--	--	--	--	--	--	-0.70	0.000
Child 0-4 yrs in HH	--	--	--	--	--	--	--	--	1.11	0.000
Result Code										
Cooperation	--	--	--	--	--	--	--	--	--	--
Refusal	-0.18	0.132	-0.12	0.048	-0.56	0.000	-0.43	0.002	-0.64	0.000
Noncontact	-0.62	0.000	-0.12	0.301	--	--	--	--	--	--
Area										
London	-0.50	0.000	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
North East	n.s.	n.s.	0.06	0.757	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
North West	n.s.	n.s.	0.13	0.367	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Yorkshire	n.s.	n.s.	0.37	0.014	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
East Midlands	n.s.	n.s.	-0.08	0.587	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
West Midlands	n.s.	n.s.	-0.10	0.503	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
East of England	n.s.	n.s.	0.12	0.379	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
London	n.s.	n.s.	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
South East	n.s.	n.s.	0.20	0.112	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Table F1. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
South West	n.s.	n.s.	0.35	0.020	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Wales	n.s.	n.s.	0.10	0.546	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Scotland	n.s.	n.s.	-0.54	0.000	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Household Char.										
House	--	--	--	--	--	--	--	--	n.s.	n.s.
Flat	--	--	0.14	0.047	-0.17	0.064	-0.44	0.013	n.s.	n.s.
Caravan, Other	--	--	2.15	0.035	0.44	0.547	--	--	n.s.	n.s.
Own	0.53	0.000	-0.30	0.002	0.76	0.000	0.91	0.000	n.s.	n.s.
Rooms	0.08	0.040	0.23	0.000	-0.08	0.001	n.s.	n.s.	n.s.	n.s.
Council House	0.29	0.038	--	--	0.23	0.035	0.61	0.001	n.s.	n.s.
Lowest Floor 1 or 2	-0.73	0.000	n.s.	n.s.	n.s.	n.s.	-0.42	0.043	n.s.	n.s.
Working Adult	0.32	0.003	-0.22	0.001	--	--	n.s.	n.s.	n.s.	n.s.
0 Cars	n.s.	n.s.	--	--	--	--	n.s.	n.s.	n.s.	n.s.
1 Car	n.s.	n.s.	0.08	0.193	-0.25	0.002	n.s.	n.s.	n.s.	n.s.
2 Cars	n.s.	n.s.	0.68	0.000	0.06	0.602	n.s.	n.s.	n.s.	n.s.
3+ Cars	n.s.	n.s.	0.63	0.000	0.19	0.339	n.s.	n.s.	n.s.	n.s.
1 Adult	--	--	--	--	--	--	n.s.	n.s.	--	--
2 Adults	0.15	0.164	-0.13	0.027	-0.20	0.018	n.s.	n.s.	0.04	0.679
3 Adults	-0.06	0.764	-0.49	0.000	-0.62	0.000	n.s.	n.s.	-0.70	0.000
4+ Adults	-0.59	0.014	-0.85	0.000	-0.14	0.453	n.s.	n.s.	-1.22	0.000
No Children	n.s.	n.s.	--	--	--	--	--	--	--	--
1 Child	n.s.	n.s.	-0.08	0.300	0.003	0.971	-0.28	0.075	--	--
2 Children	n.s.	n.s.	-0.21	0.009	0.24	0.030	0.15	0.405	--	--
3+ Children	n.s.	n.s.	-0.50	0.000	0.12	0.400	0.77	0.003	--	--
All White	n.s.	n.s.	n.s.	n.s.	--	--	--	--	--	--
Mixed Race Only	n.s.	n.s.	n.s.	n.s.	-0.71	0.224	--	--	-1.19	0.085
Asian Only	n.s.	n.s.	n.s.	n.s.	-0.31	0.158	--	--	0.20	0.444
Black Only	n.s.	n.s.	n.s.	n.s.	-0.66	0.002	--	--	-0.71	0.004
Chinese/Other Only	n.s.	n.s.	n.s.	n.s.	-1.06	0.001	--	--	-1.38	0.000
Mixed HH	n.s.	n.s.	n.s.	n.s.	-0.61	0.000	--	--	0.21	0.297

Table F1. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
Total Calls to HH	n.s.	n.s.	n.s.	n.s.	-0.03	0.044	n.s.	n.s.	n.s.	n.s.
Deprivation Indicators										
Employment	n.s.	n.s.	n.s.	n.s.	-0.96	0.000	-0.29	0.049	-0.30	0.002
Education	n.s.	n.s.	-0.31	0.000	n.s.	n.s.	0.34	0.009	0.47	0.000
Health/Disability	n.s.	n.s.	-0.21	0.000	0.20	0.006	n.s.	n.s.	n.s.	n.s.
Housing	n.s.	n.s.	n.s.	n.s.	-0.33	0.000	n.s.	n.s.	n.s.	n.s.
Interviewer Char.										
<39 years old	0.41	0.077	n.s.	n.s.	0.32	0.049	n.s.	n.s.	n.s.	n.s.
40- 49 years old	0.47	0.009	n.s.	n.s.	-0.04	0.699	n.s.	n.s.	n.s.	n.s.
50-59 years old	0.28	0.032	n.s.	n.s.	0.14	0.093	n.s.	n.s.	n.s.	n.s.
60+ years old	--	--	n.s.	n.s.	--	--	n.s.	n.s.	n.s.	n.s.
<1 year experience	-0.26	0.151	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.33	0.013
1-2 years experience	-0.30	0.046	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.03	0.789
3-8 years experience	0.02	0.887	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.16	0.120
9+ years experience	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Interviewer Survey										
Confident	n.s.	n.s.	n.s.	n.s.	-0.15	0.025	n.s.	n.s.	n.s.	n.s.
Keep Conversation Going- Str Disagree	n.s.	n.s.	0.54	0.023	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Refusal Affects Behavior:										
Always/Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.15	0.690	n.s.	n.s.
Sometimes/Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.35	0.011	n.s.	n.s.
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--	n.s.	n.s.
Refusal Affects Behavior:										
Always	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.36	0.565
Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	-0.65	0.009
Sometimes	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.09	0.517
Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.04	0.704
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Refusal Affects How I Feel:										
Always	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.32	0.325
Frequently	1.04	0.001	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.03	0.857

Table F1. *Continued*

	Type of Housing Unit N=17,759		Council N=17,053		Working Adult N=15,575		White N=16,724		Children N=14,910	
	Coefficient	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**	**Coefficient**	**p value**
Sometimes	0.89	0.002	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.15	0.262
Rarely	0.72	0.012	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.29	0.032
Never	0.88	0.004	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Guess Type of People in Home:										
Always	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	--	--
Frequently	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.39	0.000
Sometimes	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.24	0.033
Rarely	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.14	0.282
Never	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.54	0.033
Ask to Enter Home:										
Always	0.58	0.030	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Frequently/Sometimes/ Rarely	--	--	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Never	0.13	0.239	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Use Wide Variety of Approaches-Str Agree	-0.37	0.008	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Survey										
Expenditure & Food	--	--	--	--	--	--	--	--	--	--
Family Resources	0.26	0.188	0.19	0.064	0.04	0.741	0.15	0.500	0.02	0.892
General Household	-0.04	0.791	0.09	0.325	-0.16	0.122	0.45	0.027	-0.12	0.324
Omnibus	-0.04	0.805	0.10	0.283	-0.06	0.613	0.39	0.072	-0.24	0.051
National Travel	-0.19	0.238	-0.04	0.687	-0.17	0.111	-0.04	0.855	-0.21	0.098
Labour Force	-0.39	0.017	0.20	0.047	0.03	0.777	0.14	0.483	0.25	0.088
Final Rho (se)	0.033 (0.020)		0.076 (0.011)		0.004 (0.009)		0.062 (0.026)		0.024 (0.012)	
95% CI	[0.010, 0.103]		[0.058, 0.100]		[0.000, 0.213]		[0.027, 0.138]		[0.009, 0.062]	

n.s. = not significant for the accuracy that observation and therefore not shown in the final model

Table F2. Estimated Coefficients of the Final Cross-Classified Model Predicting Accuracy of the Council Housing Interviewer Observation (estimated using MlwiN, MCMC method) † (standard errors in parentheses)

	Council N=17,053	
	Coefficient	**(se)**
True Value		
Council House	-0.450	(0.063)
Result Code		
Cooperation	--	--
Refusal	-0.139	(0.060)
Noncontact	-0.063	(0.114)
Area		
North East	0.138	(0.238)
North West	0.180	(0.178)
Yorkshire	0.464	(0.202)
East Midlands	-0.004	(0.194)
West Midlands	-0.007	(0.192)
East of England	0.161	(0.181)
London	--	--
South East	0.259	(0.164)
South West	0.465	(0.188)
Wales	0.118	(0.209)
Scotland	-0.543	(0.182)
Household Char.		
House	--	--
Flat	0.197	(0.076)
Caravan, Other	2.692	(1.279)
Rooms	0.212	(0.021)
Working Adult	-0.310	(0.064)
0 Cars	--	--
1 Car	0.012	(0.063)
2 Cars	0.532	(0.091)
3+ Cars	0.282	(0.143)
No Children	--	--
1 Child	-0.090	(0.079)
2 Children	-0.165	(0.081)
3+ Children	-0.464	(0.104)
Deprivation Indicators		
Education	-0.271	(0.060)
Health/Disability	-0.283	(0.052)
Interviewer Survey		
Keep Conversation Going-Str Disagree	0.640	(0.233)
Survey		
Expenditure & Food	--	--
Family Resources	0.183	(0.105)
General Household	0.118	(0.093)
Omnibus	0.105	(0.092)
National Travel	-0.037	(0.096)
Labour Force	0.184	(0.097)

†The values in each cell are the point estimate (the means of 45,000 MCMC samples, with burn-in of 500). Standard errors (*se*) are calculated as the standard deviations of the estimates from the MCMC samples. DIC=deviance information criterion.

Appendices for Chapter 3

Appendix 3A: Further Details on Data Preparation Specific to the Microm Data

For the UBR sample, the following Microm categories have no households reporting high income in the survey: “slightly higher than average proportion of families with children” (Family type, category six); “well below the average proportion of foreigners” (Foreign, category four); and a “slightly positive balance” in mobility (Mobility, category six). Similarly, in the UBN sample, there are no households living in an area with a “very low proportion of foreigners” (category three) and reporting high income. When analyzing the samples separately to predict income, these Microm categories are collapsed with the next highest category.

Some data cannot be combined with another category due to the definition of the categories and have to be excluded. The category of House type indicating “mostly households combined with commercial space” cannot be combined with any of the other categories designating apartments or single family homes. This category has no households which report being on UB in the GP sample, having medium or high income in the UBR sample, or high income in the UBN sample, thereby creating empty cells in the analysis if this category is not excluded. The missing values on the Microm indicators are not evenly distributed across the categories of the survey variables and also create problematic empty cells. The missing cases are excluded in the GP sample, where none of the missing Microm cases report being on UB, and in the UBN sample, where none of the missing cases report having high income.

Appendix 3B: Details of the Model Development

Determining the optimal models for the analysis involved examining the nesting of cases within interviewers and areas, which would indicate whether cross-classified multilevel modeling is required, as has been necessary in the analyses of other face-to-face survey data (O'Muircheartaigh and Campanelli 1999; Durrant et al. 2010). Although there was some overlap of interviewers and areas, sufficient cross-classification was not evident in the data. In the analysis dataset, which included all three refreshment samples, one interviewer worked 61% of the 335 areas and 43% of the 270 interviewers only worked in one area.

Considering simpler two-level models, the area and interviewer random effects were tested separately using a minimum of two cases per interviewer or area. Since UB is dichotomous, a logistic regression using xtlogit in Stata determined that the random effects for both of these cluster variables were not significant ($X^2(1)= 0.00031$, $p= 0.493$ for area; $X^2(1)=0.66$, $p=0.209$ for interviewers) when all other variables are in the model. Repeating the test for random effects separately for each of the samples led to the same conclusion. Therefore, the analysis using auxiliary data to predict whether someone in the household is on UB uses a simple logistic regression.

The income variable has three ordered categories, necessitating the use of an ordered logit model. To study the random effects in this type of model, the cluster command was used. Comparing the significance tests of the categories and overall variables between the clustered and unclustered model revealed no notable differences in either the full model or when running the samples separately. This conclusion was confirmed by running similar ordered probit models with random effects (using reoprob; see Frechette 2001) and evaluating the significance of the random effects. Therefore, as with the UB analysis, a simpler model was run to predict income.

For the two models predicting UB and income, interaction effects between the sample and the auxiliary variables were explored to determine if the samples should be analyzed separately. Since PASS develops the nonresponse weights separately for each sample (Trappmann 2011), there is a precedent of analyzing the data this way for the current analysis. In addition, intuitively, the differences between the types of households in each sample may lead to differences in the ability of each type of auxiliary data to predict income or UB. This was confirmed by exploratory analysis.

Appendix 3C: Cross-Tabulations Showing the Accuracy of the Interviewer Observations for UB and Income, across all Samples

Table C1. Frequency of the Interviewers Observations of Unemployment Benefit Status and the Percent within Each Observational Category that Corresponds with the Self-Reported Value from the Survey, for all PASS Samples

		Unemployment benefit: Self-reported								
		General Population Refreshment (GP)			UB Refreshment, new regions (UBR)			UB Refreshment, new in the last year (UBN)		
UB: Interviewer Observed		**On UB**	**Not on UB**		**On UB**	**Not on UB**		**On UB**	**Not on UB**	
		n=85	n=1292		n=913	n=263		n=349	n=311	
	(N)	(%)	(%)	(N)	(%)	(%)	(N)	(%)	(%)	
On UB	200	27.5	72.5	719	85.9	14.1	315	71.4	28.6	
Not on UB	1146	2.4	97.6	428	63.1	36.9	332	36.1	63.9	
Missing	31	9.7	90.3	29	77.6	22.4	13	30.8	69.2	

Table C2. Frequency of the Interviewers Observations of Income and the Percent within Each Observational Category that Corresponds with the Self-Reported Value from the Survey, for all PASS Samples

		Income: Self-reported										
		General Population Refreshment (GP)			UB Refreshment, new regions (UBR)				UB Refreshment, new in the last year (UBN)			
Income: Interviewer Observed		**Low**	**Medium**	**High**		**Low**	**Medium**	**High**		**Low**	**Medium**	**High**
		n=391	n=474	n=512		n=1060	n=94	n=22		n=510	n=116	n=34
	(N)	(%)	(%)	(%)	(N)	(%)	(%)	(%)	(N)	(%)	(%)	(%)
Low	329	54.7	31.6	13.7	803	92.2	6.8	1.0	379	85.5	12.7	1.8
Medium	782	22.0	37.8	40.2	331	85.2	10.9	3.9	249	64.7	26.5	8.8
High	235	11.9	26.0	62.1	13	69.2	23.1	7.7	19	73.7	10.5	15.8
Missing	31	35.5	41.9	22.6	29	100.0	0.0	0.0	13	84.6	0.0	15.4

Appendix 3D: Distribution of the Microm Variables across Samples

Table D1. Distribution of Microm Variables Used in the Analysis, Overall and within Each Sample

Variable name	Description	Overall	General Population Refreshment (GP)	UB Refreshment, new regions (UBR)	UB Refreshment, new in the last year (UBN)
		n=3213	n=1377	n=1176	n=660
		(%)	(%)	(%)	(%)
House type	Concentration of family homes				
	(1) 1-2 family homes on streets with homogeneous building structures	14.6	23.0	6.5	11.5
	(2) 1-2 family homes on streets with heterogeneous building structures	20.1	29.6	12.8	13.5
	(3) 3-5 family homes	19.7	14.4	22.6	25.3
	(4) 6-9 family homes	19.4	12.2	25.3	23.8
	(5) Apartment block with 10-19 households	12.9	10.0	16.8	12.1
	(6) High rise buildings with 20+ households	7.5	4.9	10.0	8.7
	(7) Mostly households combined with commercial space	1.3	1.2	1.2	1.5
Mobility	Measure of households moving in and out				
	(1) Very strongly negative rate - moving out	13.3	7.5	18.7	15.9
	(2) Strongly negative rate - moving out	12.9	9.4	15.3	15.9
	(3) Negative rate - moving out	12.0	9.1	15.0	12.6
	(4) Slightly negative rate - moving out	10.6	10.2	11.3	10.3
	(5) Balanced rate - moving out	10.3	11.0	8.9	11.2
	(6) Slightly positive rate - moving in	9.0	10.8	7.2	8.5
	(7) Positive rate - moving in	9.2	11.8	7.0	7.6
	(8) Strongly positive rate - moving in	8.8	12.4	5.4	7.3
	(9) Very strongly positive rate - moving in	9.4	13.1	6.4	7.1
Under 30	Percent of people under 30 years old				
	(0) Up to 5%	11.5	15.6	8.0	9.4
	(1) 5% - 10%	7.6	9.7	5.8	6.4
	(2) 10% - 15%	8.0	9.9	6.8	6.2
	(3) 15% - 20%	9.8	11.0	9.4	7.9
	(4) 20% - 25%	9.0	8.8	8.5	10.3
	(5) 25% - 30%	9.5	8.1	10.2	11.2
	(6) 30% - 35%	8.9	8.9	8.6	9.1
	(7) 35% - 40%	6.8	5.7	7.9	7.0
	(8) 40% - 50%	12.6	9.8	15.3	13.8
	(9) Over 50%	11.8	7.8	14.7	15.1
Foreign	Proportion of foreigners				
	(1) No foreigners	10.7	11.2	8.9	12.7
	(2) Extremely low proportion	11.1	13.3	9.6	9.1
	(3) Very low	8.6	9.5	7.9	7.9
	(4) Well below average	9.3	13.4	5.9	7.0

Table D1. *Continued*

Variable name	Description	Overall	General Population Refreshment (GP)	UB Refreshment, new regions (UBR)	UB Refreshment, new in the last year (UBN)
		n=3213	n=1377	n=1176	n=660
		(%)	(%)	(%)	(%)
	(5) Below average	9.0	10.8	7.3	8.2
	(6) Slightly below average	10.0	9.5	10.9	9.5
	(7) Average	11.8	10.6	13.4	11.5
	(8) Above average	11.1	7.8	13.4	13.8
	(9) Highest proportion	13.9	9.2	17.9	16.7
Family type	Composition of families				
	(1) Mostly single person households	9.0	5.2	10.8	13.8
	(2) Well above average proportion of single person households	13.6	8.9	19.6	12.4
	(3) Above average proportion of single person households	12.2	8.9	15.7	13.0
	(4) Slightly higher than average proportion of single person households	11.6	9.2	14.5	11.5
	(5) Mixed family structure	11.8	10.5	11.8	14.4
	(6) Slightly higher than average proportion of families with children	10.1	11.0	8.6	10.8
	(7) Above average proportion of families with children	9.2	12.3	6.4	7.9
	(8) Well above average proportion of families with children	9.8	15.2	4.6	7.7
	(9) Almost exclusively families with children	8.2	14.1	3.2	4.9
Status	Status (wealth & prominence) of households				
	(1) Lowest status	20.4	10.2	33.1	19.1
	(2) Very low status	13.6	9.2	16.8	17.0
	(3) Well below average status	13.1	13.0	12.7	14.1
	(4) Below average status	9.8	10.5	8.8	10.0
	(5) Slightly below average status	8.8	11.1	6.5	8.2
	(6) Average status	9.8	12.0	7.1	9.8
	(7) Slightly above average status	7.8	10.2	4.9	7.9
	(8) Above average status	7.2	10.8	3.5	6.2
	(9) Highest status	5.0	8.3	1.8	4.1
Missing		4.5	4.7	4.8	3.6

Appendix 3E: Pseudo R^2 values for the final models predicting UB and income

Table E1. Pseudo R^2 Values and Model Significance for Models Predicting Self-Reported UB and Income, Comparing Each Auxiliary Data Source Separately and Combined

		Unemployment Benefit (UB)			Income		
Sample	**Model**	**n**	**Pseudo R^2**	**Chi^2 test**	**n**	**Pseudo R^2**	**Chi^2 test**
GP		1295			1377		
	Obs. only		0.2275	p=0.0000		0.0735	p=0.0000
	Microm only		0.3031	p=0.0000		0.0617	p=0.0000
	Both		0.4368	p=0.0000		0.1054	p=0.0000
UBR		1176			1134		
	Obs. only		0.0735	p=0.0000		0.0369	p=0.0000
	Microm only		0.0465	p=0.1998		0.0844	p=0.0073
	Both		0.1121	p=0.0000		0.1030	p=0.0004
UBN		660			614		
	Obs. only		0.1041	p=0.0000		0.0574	p=0.0000
	Microm only		0.0649	p=0.1733		0.0677	p=0.1920
	Both		0.1590	p=0.0000		0.1169	p=0.0001

Appendix 3F: Cross Validation Results

Table F1. Results of the Cross Validation for all Three Samples, Predicting UB

		General Population Refreshment (GP)					UB Refreshment, new regions (UBR)					UB Refreshment, new in the last year (UBN)				
Sub sample	**Model**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**
1	**Obs**	259	0.041	0.008	p=0.487	p=0.535	236	0.157	0.013	p=0.092	p=0.000	132	0.215	0.015	p=0.001	p=0.002
	Both		0.047	0.010		p=0.875		0.168	0.014		p=0.004		0.261	0.021		p=0.285
	Microm		0.046	0.009				0.189	0.016				0.279	0.016		
2	**Obs**	213	0.088	0.016	p=0.198	p=0.019	235	0.159	0.015	p=0.243	p=0.536	132	0.255	0.018	p=0.863	p=0.928
	Both		0.088	0.016		p=0.074		0.166	0.016		p=0.859		0.257	0.020		p=0.999
	Microm		0.101	0.018				0.165	0.015				0.257	0.015		
3	**Obs**	259	0.058	0.011	p=0.190	p=0.114	235	0.166	0.015	p=0.0002	p=0.000	132	0.227	0.015	p=0.030	p=0.004
	Both		0.050	0.011		p=0.004		0.186	0.017		p=0.002		0.257	0.020		p=0.040
	Microm		0.067	0.013				0.206	0.018				0.289	0.017		
4	**Obs**	259	0.046	0.010	p=0.047	p=0.401	235	0.163	0.015	p=0.067	p=0.148	132	0.199	0.013	p=0.099	p=0.000
	Both		0.036	0.009		p=0.388		0.174	0.017		p=0.781		0.219	0.016		p=0.000
	Microm		0.040	0.009				0.176	0.016				0.273	0.015		
5	**Obs**	236	0.045	0.010	p=0.059	p=0.078	235	0.163	0.015	p=0.236	p=0.093	132	0.214	0.015	p=0.036	p=0.001
	Both		0.059	0.012		p=0.820		0.171	0.016		p=0.345		0.244	0.021		p=0.020
	Microm		0.057	0.012				0.178	0.016				0.278	0.017		

Table F2. Results of the Cross Validation for all Three Samples, Predicting Income

		General Population Refreshment (GP)					UB Refreshment, new regions (UBR)					UB Refreshment, new in the last year (UBN)				
Sub sample	**Model**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**	**n**	**Mean Difference**	**se**	**t-test with Both**	**t-test with Microm**
1	**Obs**	276	0.386	0.010	p=0.442	p=0.058	227	0.110	0.018	p=0.665	p=0.804	123	0.208	0.024	p=0.564	p=0.240
	Both		0.381	0.011		p=0.001		0.112	0.017		p=0.824		0.213	0.025		p=0.244
	Microm		0.407	0.009				0.112	0.017				0.222	0.027		
2	**Obs**	276	0.393	0.010	p=0.921	p=0.019	227	0.089	0.016	p=0.040	p=0.019	123	0.174	0.024	p=0.263	p=0.470
	Both		0.393	0.011		p=0.001		0.097	0.016		p=0.398		0.184	0.025		p=0.745
	Microm		0.419	0.009				0.098	0.016				0.181	0.025		
3	**Obs**	275	0.392	0.010	p=0.855	p=0.006	227	0.086	0.016	p=0.030	p=0.008	123	0.193	0.025	p=0.283	p=0.076
	Both		0.393	0.012		p=0.000		0.092	0.017		p=0.181		0.204	0.025		p=0.217
	Microm		0.426	0.011				0.094	0.017				0.214	0.025		
4	**Obs**	276	0.393	0.010	p=0.692	p=0.044	227	0.082	0.015	p=0.851	p=0.955	123	0.197	0.024	p=0.004	p=0.014
	Both		0.390	0.011		p=0.001		0.082	0.016		p=0.647		0.230	0.028		p=0.730
	Microm		0.418	0.010				0.081	0.015				0.227	0.029		
5	**Obs**	274	0.397	0.010	p=0.406	p=0.333	226	0.124	0.019	p=0.042	p=0.087	122	0.251	0.028	p=0.243	p=0.234
	Both		0.391	0.012		p=0.034		0.132	0.019		p=0.851		0.265	0.029		p=0.686
	Microm		0.408	0.011				0.132	0.019				0.269	0.030		

Appendices for Chapter 4

Appendix 4A. Deciding How Best to Handle the Missing Likelihood Rating for the First Contact

Once the likelihood ratings were lagged forward to the next contact, there was not a rating on the first contact. Two solutions were to impute values of the ratings for the first contact and a third solution was to drop the first contact from the analysis. The three solutions are explained and compared below. When comparing the methods, a discrete time response propensity model predicting cooperation, conditional on contact, with all of the covariates used in the final models was run for each method of handling the missing data on first contact. Any fluctuations in the model parameters were noted since this may indicate that the imputed values are strongly influencing the model. In addition, model fit statistics were compared to determine which of the options had the best fit and discrimination and if the imputation was influencing the quality of the model.

Two imputation methods

In an effort to retain the first contact (and the cases that only have one contact) in the analysis a value for the missing rating was imputed. The first imputation inserted the response rate on the first contact. This value, 26, was calculated by taking the number of cases that completed an interview on the first contact (505) and dividing it by the total number of cases in the dataset (1943). The resulting distribution of the likelihood ratings is shown in figure A1. The imputation results in a notable spike at 26, with a frequency that is three times that of the next highest likelihood rating.

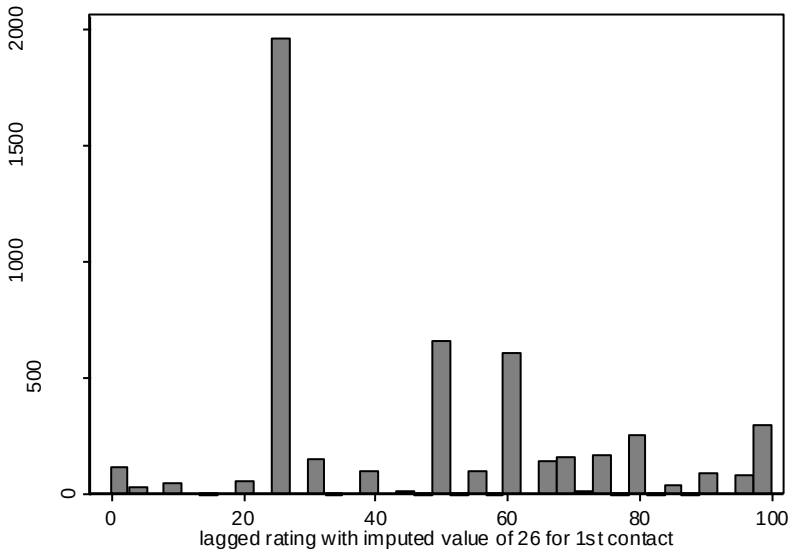


Figure A1. Distribution of the Likelihood Rating with the Value of 26 Imputed for the First Contact

A second imputation took advantage of the fact that this study is the second wave of a data collection and used the call record data from wave 1 to calculate the response probabilities at the end of data collection for each case with contact. The data were aggregated to the case level and the response probabilities were estimated from a logistic regression predicting cooperation, conditional on contact. The covariates in the model included: number of calls to the case, if the case ever had an appointment and the total number of appointments, if the

same interviewer made all the contacts to the case, if the original number provided on the sampling frame was not correct, and total number of contacts with the selected target person. The distribution of the predicted probabilities is shown below in figure A1. The figure shows a noticeable spike around 100 which is still prominent when combined with the likelihood ratings on the other contacts (see figure A2). The cases in wave 2 are the respondents to wave 1 so it is not surprising that the estimated probabilities are high.

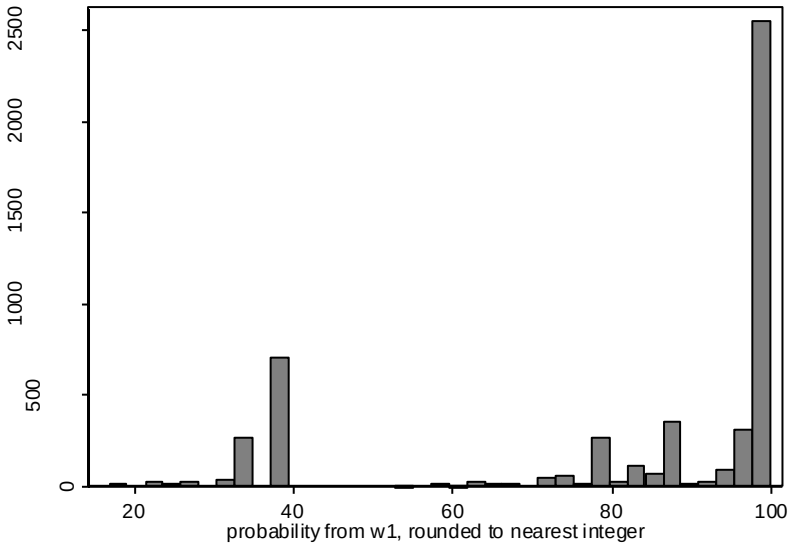


Figure A2. Distribution of the Wave 1 Probabilities for the Cases in Wave 2

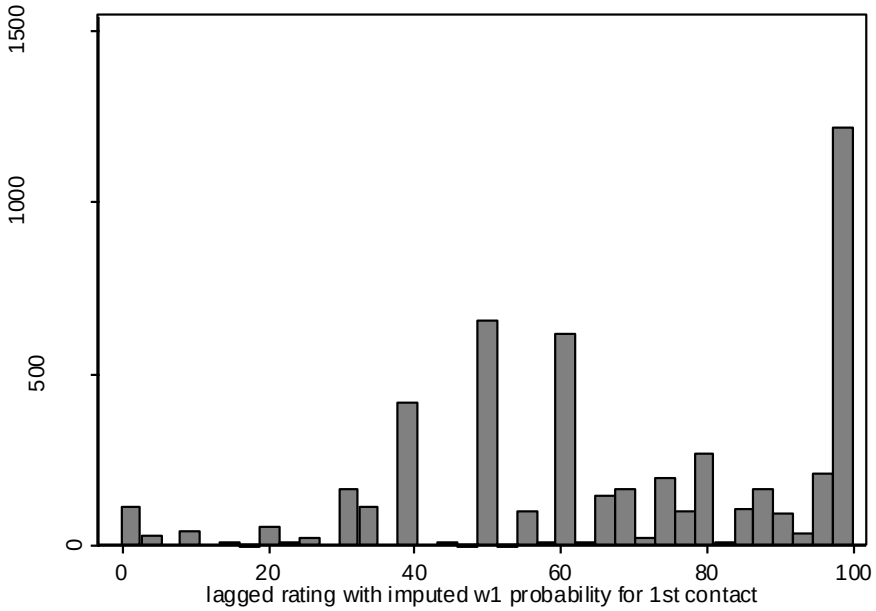


Figure A3. Distribution of the Likelihood Rating with the Wave 1 Probability Imputed for the First Contact

Dropping the first contact

An alternative to imputing the rating for the first contact is to drop the first contact from the analysis. This is an especially viable option if the imputations for the first contact result in unlikely values which interfere with the ability of the models to accurately predict cooperation.

Comparing the different methods of handling the missing rating

To test if the imputations are harmful to the models, the parameter estimates from a model without the first contact in the data were compared to the parameter estimates for each imputation noted above. Comparisons found that the imputation of the response rate (26) at the first contact did not noticeably affect the parameter estimates compared to dropping the first contact but changes to the estimates were noticeable when the propensity score was used for the imputation. The fit and discrimination statistics showed noticeable improvements when the imputation (and consequently, the first contact) was removed from the model (see table A1). The pseudo R-squared and area under the ROC curve were highest for the model without any imputation. Therefore, the best solution to the missing data for the first contact was to drop the first contact from the analysis.

Table A1. Fit and Discrimination Statistics for Three Propensity Models Applying Three Different Solutions to Handle Missing Rating on First Contact

Model	N	Pseudo R-squared	AIC	AIC df	Area under ROC Curve
Imputed 26	5034	0.0607	5501	28	0.6600
Imputed W1 Probability	5034	0.0575	5520	28	0.6601
Dropping First Contact	3091	0.0997	3268	27	0.7144

Appendix 4B. Number of Cases Available for Each Contact Number

Table B1. Number of Cases Available for Each Contact Number in the Analysis

Contact Number	Number of Cases in Analysis
2	1295
3	720
4	421
5	260
6	158
7	84
8	52
9	35
10	25
11	18
12	14
13	9

Appendix 4C. Variables Used in Analysis

Table C1. Names and Descriptions of Variables Tested in the Models

Variable	Description
Week	week of data collection
Mobile phone	call made to mobile phone
Wkday eve	call made Monday to Friday, at or after 6pm
Weekend	call made on Saturday or Sunday
Num prev calls	number of calls made to case before current call
Num prev contacts	number of contacts with case, not including current contact
Days since last contact	number of days since last contact; can be 0
Refused previously	refused on any prior call
Refused on prior contact	refused on contact immediately prior to current contact
Appt prior contact	appointment on contact immediately prior to current contact
Appt prior call	appointment on call immediately prior to current contact
Appt prior cont*Appt num	interaction between appointment on immediately prior contact and appointment number (categorized or continuous)
Appt number	number of appointments case made up to current contact
1 previous appt	1 appointment prior to current call
2-3 prev appts	2 – 3 appointments prior to current call
4+ prev appts	4 or more appointments prior to current call
NC prior call	noncontact on call immediately prior to current contact
Num prior NC	number of prior calls without contact
Cont prior call	contact on call immediately prior to current contact
Gen cont prior contact	general contact (no appointment) on contact immediately prior to current contact
Gen cont prior call	general contact (no appointment) on call immediately prior to current contact
Target person reacher, prior contact	selected respondent reached on a prior contact
Rating (cont)	likelihood rating recorded at end of previous contact
Rating #-#	likelihood rating recorded at end of previous contact, categorized
Avg rating, start of call	average of ratings of prior contacts, not including rating given at end of current contact
Int young	interviewer born before 1980, not a student
Int experience	total months worked as a telephone interviewer
Int hrs per wk	interviewer's working hours per week at present
Int prev contact	interviewer made contact with this case at least once prior to current contact
Int prior contact	same interviewer for current and immediately prior contact

Appendix 4D. Discussion of Random Effects

The discrete time hazard models in the analysis incorporate a random effect for the current interviewer. In addition, the exploration of the effect of the previous (i.e. lagged forward) interviewer on the prediction of cooperation through the ratings is explored using both random effects and fixed effects (see Appendix 4E). The use of both of these techniques involves some assumptions which may or may not hold for these data.

The use of a random intercept accounts for the heterogeneity among interviewers by allowing the intercept to vary by interviewer. The assumption is that the random effect is normally distributed, u_{0j} ~N(0,σ^2) and that the interviewers in the analysis are a random subset of a super population of interviewers. Essentially, the individual interviewers are not of interest and through training and management, interviewers are seen as interchangeable. This has the added benefit of allowing the analysis to be broadly applicable to that super population of interviewers, providing useful conclusions for all interviewers within the data collection agency and at other agencies.

When applying fixed effects to account for the influence of interviewers (by introducing a dummy variable for each interviewer), the number of covariates in the model increases by the number of interviewers minus one. This can lead to overfitting problems. Even if these are avoided, the conclusions of the analysis are specific to the (in this case, twenty-two) interviewers in the analysis. However, fixed effects overcome the assumption that the random effect of the interviewers is independent of the covariates in the model.

Tutz and Oelker (2014) provide an argument for the use of fixed effects over random effects, specifically when combined with a regularization method that clusters interviewers that are the same. This approach will overcome the underlying assumption of the normal distribution that all interviewers differ with respect to cooperation, and none are the same. This is a valuable approach but the use of fixed effects in the analysis of survey data is not widely accepted because survey agencies want the results to be applicable to all interviewers (i.e. the super population of interviewers mentioned above). Moving the field of survey methodology toward more accurate models for modeling interviewer effects is an agenda item for future analyses.

Appendix 4E. Results of Tests of the Random Effect of the Lagged Interviewer on the Lagged Rating

Three different approaches to modeling the interviewer effect on the likelihood ratings were applied: fixed effects for the lagged interviewer (corresponding the lagged rating), interactions between the lagged interviewers and the lagged ratings, and random coefficients for the ratings accounting for the effect of lagged interviewers.

Tests of the Ratings-only model, including the random effect for the current interviewer, found that when the lagged interviewer identification numbers were included in the model as fixed effects, the coefficients of four interviewers (ID numbers 3, 7, 18, and 22) were significant (at either α=0.05 or 0.10) relative to the reference interviewer (number 12). Of these significant effects, only two remained significant at α=0.10 when all covariates were entered into the model (see table E1). Overall, this analysis does not support significant variation in the way the interviewers use of the ratings to predict cooperation.

The investigation continued by introducing interactions between the ratings and the interviewer who made the rating. Again, starting with the Ratings-only model and retaining the random effect for the current interviewer, the model found few significant interaction effects (see table E2). However, the main effects of the interviewers corresponding to the significant interactions are large (note that the continuous version of the likelihood rating had to be used in this test). Once all covariates are included in the model, only one interaction is significant at the α=0.05 level (interviewer 3) and two others are marginally significant. This investigation does not support a significant interviewer effect on the likelihood rating.

Table E1. Discrete Time Hazard Propensity Models Predicting Cooperation, Conditional on Contact, with Random Intercept for Current Interviewer and Fixed Effects for Lagged Interviewer; Parameters shown as Odds Ratios with p-values

	Odds Ratio	p-value
	N = 3091	
(contact numbers not shown)		
Week	1.02	0.477
Mobile phone	1.45	0.001
Wkday eve	0.68	0.000
Weekend	0.76	0.039
Num prev calls	0.97	0.005
Days since last contact	0.96	0.001
Refused previously	0.62	0.069
Refused on prior contact	0.44	0.081
No contact, prior call	0.68	0.000
1 previous appt	1.48	0.061
2-3 prev appts	2.28	0.005
4+ prev appts	3.27	0.004
Target personreached, prior contact	1.21	0.068
Interviewer 1 (lagged)	1.41	0.259
Interviewer 2 (lagged)	1.07	0.819
Interviewer 3 (lagged)	2.06	0.242
Interviewer 4 (lagged)	1.21	0.474
Interviewer 5 (lagged)	0.33	0.755
Interviewer 6 (lagged)	1.49	0.408
Interviewer 7 (lagged)	1.66	0.074
Interviewer 8 (lagged)	1.19	0.548
Interviewer 9 (lagged)	0.74	0.533
Interviewer 10 (lagged)	1.01	0.975
Interviewer 11 (lagged)	1.24	0.407
Interviewer 13 (lagged)	1.02	0.948
Interviewer 14 (lagged)	0.96	0.875
Interviewer 15 (lagged)	1.74	0.665
Interviewer 16 (lagged)	0.85	0.660
Interviewer 17 (lagged)	1.38	0.260
Interviewer 18 (lagged)	2.24	0.017
Interviewer 20 (lagged)	1.17	0.570
Interviewer 21 (lagged)	1.14	0.671
Interviewer 22 (lagged)	1.54	0.114
Rating 10-19	0.21	0.033
Rating 20-29	0.31	0.051
Rating 30-39	0.31	0.020
Rating 40-49	0.39	0.060
Rating 50	0.40	0.043
Rating 51-59	0.37	0.057
Rating 60-69	0.38	0.040
Rating 70-79	0.50	0.153
Rating 80-89	0.79	0.628
Rating 90-95	0.87	0.798
Rating 96-100	0.95	0.918
Avg rating, start of call	1.00	0.378
Current interviewer effect:		
rho	0.014	0.001
se	0.008	

Table E2. Discrete Time Hazard Propensity Models Predicting Cooperation, Conditional on Contact, with Random Intercept for Current Interviewer and Interaction Effects between the Lagged Ratings and the Lagged Interviewer; Parameters shown as Odds Ratios with p-values

	Odds Ratio	**p-value**
	N = 3091	
(contact numbers not shown)		
Interviewer 1 (lagged)	0.14	0.156
Interviewer 2 (lagged)	2.84	0.330
Interviewer 3 (lagged)	63.42	0.001
Interviewer 4 (lagged)	3.60	0.209
Interviewer 5 (lagged)	0.75	0.862
Interviewer 6 (lagged)	4.77	0.224
Interviewer 7 (lagged)	1.07	0.953
Interviewer 8 (lagged)	0.59	0.682
Interviewer 9 (lagged)	23.04	0.063
Interviewer 10 (lagged)	3.50	0.267
Interviewer 11 (lagged)	2.11	0.456
Interviewer 13 (lagged)	2.75	0.349
Interviewer 14 (lagged)	3.40	0.243
Interviewer 15 (lagged)	13.02	0.359
Interviewer 16 (lagged)	1.37	0.805
Interviewer 17 (lagged)	2.56	0.399
Interviewer 18 (lagged)	10.35	0.044
Interviewer 20 (lagged)	0.35	0.367
Interviewer 21 (lagged)	2.23	0.431
Interviewer 22 (lagged)	4.05	0.183
Int 1 (lag)*rating	1.04	0.065
Int 2 (lag)*rating	0.99	0.344
Int 3 (lag)*rating	0.89	0.029
Int 4 (lag)*rating	0.98	0.257
Int 5 (lag)*rating	1.00	0.962
Int 6 (lag)*rating	0.99	0.551
Int 7 (lag)*rating	1.01	0.554
Int 8 (lag)*rating	1.01	0.538
Int 9 (lag)*rating	0.96	0.046
Int 10 (lag)*rating	0.99	0.361
Int 11 (lag)*rating	0.99	0.659
Int 13 (lag)*rating	0.99	0.465
Int 14 (lag)*rating	0.98	0.213
Int 15 (lag)*rating	0.96	0.445
Int 16 (lag)*rating	0.99	0.757
Int 17 (lag)*rating	0.99	0.522
Int 18 (lag)*rating	0.98	0.195
Int 20 (lag)*rating	1.02	0.246
Int 21 (lag)*rating	0.99	0.734
Int 22 (lag)*rating	0.99	0.448
Rating (continuous)	1.03	0.028
Avg rating, start of call	1.01	0.084
Current interviewer effect:		
rho	0.020	0.000
se	0.009	

The final test involved including random coefficients for the ratings accounting for the effect of the interviewer who made the ratings. As before, the Ratings-only model with a random intercept for the current interviewer was used to examine the significance of the random coefficients but the average likelihood rating variable was excluded from the model. As is consistent with the other tests, the contribution to the variance due to the random coefficients was not significant (variance =0.0096, se=0.0706). See table E3 below. When the average likelihood rating variable or the other covariates are included in the model, the variance of the coefficients is essentially zero (e.g., when just the average likelihood rating variable is added (not the covariates in the full model), the variance of the coefficients is 6.59e-08, se=0.0011).

Table E3. Discrete Time Hazard Propensity Models Predicting Cooperation, Conditional on Contact, with Random Intercept for Current Interviewer and Random Coefficients for the Ratings Accounting for the Interviewer Who Made the Rating; Parameters shown as Coefficients with p-values

	Coefficient	**p-value**
	N = 3091	
(contact numbers not shown)		
Rating 10-19	-0.61	0.352
Rating 20-29	-0.21	0.686
Rating 30-39	0.39	0.272
Rating 40-49	0.66	0.068
Rating 50	0.70	0.015
Rating 51-59	0.66	0.070
Rating 60-69	0.81	0.005
Rating 70-79	1.03	0.001
Rating 80-89	1.64	0.000
Rating 90-95	1.78	0.000
Rating 96-100	1.78	0.000
Current interviewer effect:		
variance	0.062	
se	0.030	
Lagged interviewer effect:		
variance	0.0096	
se	0.0706	

Appendix 4F. Formulas for the Calculation of the Net Reclassification Index (NRI)

Below are the formulas for the calculation of the Net Reclassification Index (NRI) adapted from Pencina et al. (2008) to specifically characterize the data in this analysis. Movement "up" means moving from a low or medium probability tertile to a category with a higher probability of cooperation (medium or high). Movement "down" is the opposite.

The four probabilities of movement:

$$\hat{P}(up|Respondent) = \hat{p}_{up,R} = \frac{\#\ cooperative\ contacts\ moving\ up}{\#\ cooperative\ contacts}$$

$$\hat{P}(down|Respondent) = \hat{p}_{down,R} = \frac{\#\ cooperative\ contacts\ moving\ down}{\#\ cooperative\ contacts}$$

$$\hat{P}(up|Nonrespondent) = \hat{p}_{up,NR} = \frac{\#\ noncooperative\ contacts\ moving\ up}{\#\ noncooperative\ contacts}$$

$$\hat{P}(down|Nonrespondent) = \hat{p}_{down,NR} = \frac{\#\ noncooperative\ contacts\ moving\ down}{\#\ noncooperative\ contacts}$$

The NRI calculation:

$$\widehat{\boldsymbol{NRI}} = \left(\hat{p}_{up,R} - \hat{p}_{down,R}\right) - \left(\hat{p}_{up,NR} - \hat{p}_{down,NR}\right)$$

Appendix 4G. Data Collection Progress by Date and the Selection of Monitoring Dates for Responsive Survey Design "Daily" Models

Table G1. Data Collection Progress by Date

Date	Notes	Day of data collection	Number of contacts made	Contact numbers	Contact number with highest freq	Frequency for highest	Contact number with second highest freq	Frequency for second highest
29 Oct	Monday	1	159	1-2	1	92%	2	8%
30 Oct		2	325	1-3	1	76%	2	22%
31 Oct		3	246	1-5	1	70%	2	20%
1 Nov	Holiday							
2 Nov		4	292	1-6	1	58%	2	32%
3 Nov		5	187	1-4, 6, 7	1	57%	2	30%
4 Nov	Sunday	6	10	2-4	2	50%	3	40%
5 Nov		7	216	1-6, 8	2	46%	3	21%
6 Nov		8	179	1-8	3	31%	2	27%
7 Nov		9	124	1-7	2	35%	1	19%
8 Nov		10	79	1-7, 9, 10	2	32%	1	23%
9 Nov		11	103	1-10	2	41%	3	24%
10 Nov		12	46	1-6, 8, 11	4	33%	3	26%
11 Nov	Sunday	13	2	5, 7	5	50%	7	50%
12 Nov		14	74	2-7, 9	2	24%	3	24%
13 Nov		15	56	2-8, 12	4	25%	3	23%
14 Nov		16	49	2-9, 13	3	27%	2	18%
15 Nov		17	32	2-11	5	22%	2, 3	16%
16 Nov		18	26	2-7, 11	4	31%	3	23%
17 Nov		19	16	2-7, 9,12	5	25%	4	31%
18 Nov	Sunday - no calls							
19 Nov	**Sample release**	**20**	**304**	**1-9, 11-13**	**1**	**75%**	**2**	**14%**
20 Nov		21	305	1-7	1	66%	2	25%
21 Nov		22	233	1-6, 8, 10	1	55%	2	28%
22 Nov		23	222	1-11	1	46%	2	30%
23 Nov		24	269	1-8, 11, 12	2	39%	1	24%
24 Nov		25	171	1-7, 9-11,13	1	48%	2	27%
25 Nov	Sunday - no calls							
26 Nov		26	140	1-9, 12	2	29%	3	25%
27 Nov		27	156	1-10	2	32%	3	26%
28 Nov		28	56	1-5, 6, 9	3	30%	4	25%
29 Nov		29	106	1-8	2	28%	3	25%
30 Nov		30	109	1-7, 9,10,13	2	26%	3	26%
1 Dec		31	52	1-7, 10,11,13	2	25%	3	19%
2 Dec	Sunday	32	3	3, 5, 6	1	34%	2, 3	33%
3 Dec		33	39	1-6, ,8, 9	3	23%	5	21%
4 Dec		34	57	1-7, 9-12	4	28%	2, 3	23%
5 Dec		35	30	1-8, 11	4	27%	3	20%
6 Dec		36	46	1-11	6	24%	3, 4	17%
7 Dec		37	35	1-6, 9, 12	4	26%	3	20%
8 Dec		38	174	1-12	1	23%	2	21%
9 Dec	Sunday - 3 calls	39						
10 Dec		40	50	1-10, 12, 13	3	18%	5	16%
11 Dec		41	40	1-12	6	20%	2, 5	18%
12 Dec		42	55	1-8, 10,12,13	4	27%	5	20%
13 Dec		43	105	1-11, 13	2	26%	1	17%
14 Dec		44	56	1-12	4	16%	5, 6	16%

Table G1 shows the progress of the data collection by day in terms of number of new contacts made and the range of contact numbers in the data for each date. This table was used to decide which days of the 44 possible to run the "daily" response propensity models. Daily models were not run on Sundays and the holiday when the number of calls made is low or none, and the day before the new sample release. The first date that sufficient data were available to successfully run the hazard models was November 5. After this date, at least two days were chosen per week, varying the day of the week that the models were run. These dates were: November 12, 17, 20, 21, 28, and 29, and December 3, 7, 8, 12, and 13. The final day of data collection was December 14.

Appendix 4H. Full Model Used in Analyses

Table H1. Four Versions of the Discrete Time Hazard Propensity Models Predicting Cooperation, Conditional on Contact, with Random Effects for Interviewers and All Contact Numbers; Parameters shown as Odds Ratios with p-values (in parenthesis)

	Empty		**Classic**		**Ratings-Only**		**Classic+**	
	N = 3091		N = 3091		N = 3091		N = 3091	
contact2	0.480 p=0.000	***	0.460 p=0.001	***	0.137 p=0.000	***	0.752 p=0.574	
contact3	0.389 p=0.000	***	0.258 p=0.000	***	0.121 p=0.000	***	0.469 p=0.159	
contact4	0.259 p=0.000	***	0.164 p=0.000	***	0.0855 p=0.000	***	0.308 p=0.036	*
contact5	0.286 p=0.000	***	0.154 p=0.000	***	0.0938 p=0.000	***	0.289 p=0.040	*
contact6	0.267 p=0.000	***	0.137 p=0.000	***	0.0924 p=0.000	***	0.258 p=0.036	*
contact7	0.231 p=0.000	***	0.120 p=0.000	***	0.0763 p=0.000	***	0.211 p=0.025	*
contact8	0.189 p=0.000	***	0.0859 p=0.000	***	0.0643 p=0.000	***	0.161 p=0.014	*
contact9	0.177 p=0.000	***	0.101 p=0.001	***	0.0627 p=0.000	***	0.195 p=0.045	*
contact10	0.0901 p=0.001	**	0.0578 p=0.001	**	0.0340 p=0.000	***	0.113 p=0.030	*
contact11	0.0615 p=0.007	**	0.0297 p=0.002	**	0.0223 p=0.000	***	0.0616 p=0.024	*
contact12	0.173 p=0.023	*	0.0950 p=0.010	*	0.0658 p=0.001	**	0.194 p=0.111	
contact13	0.294 p=0.130		0.169 p=0.062		0.116 p=0.013	*	0.376 p=0.358	
Week			0.976 p=0.400				1.033 p=0.318	
Mobile phone			1.518 p=0.000	***			1.466 p=0.001	***
Wkday eve			0.705 p=0.001	***			0.689 p=0.000	***
Weekend			0.804 p=0.092				0.764 p=0.042	*
Num prev calls			0.974 p=0.024	*			0.967 p=0.006	**
Days since last contact			0.953 p=0.000	***			0.956 p=0.000	***
Refused previously			0.636 p=0.056				0.625 p=0.061	
Refused on prior contact			0.557 p=0.105				0.442 p=0.071	

Table H1. *Continued*

	Empty	Classic	Ratings-Only	Classic+
	N = 3091	N = 3091	N = 3091	N = 3091
No contact, prior call		0.686 *** p=0.000		0.679 *** p=0.000
1 previous appt		1.832 ** p=0.002		1.543 * p=0.034
2-3 prev appts		3.107 *** p=0.000		2.431 ** p=0.002
4+ prev appts		4.550 *** p=0.000		3.557 ** p=0.001
Target personreached, prior contact		1.535 *** p=0.000		1.202 p=0.064
Rating 10-19			0.485 p=0.273	0.226 * p=0.036
Rating 20-29			0.714 p=0.511	0.303 * p=0.040
Rating 30-39			1.166 p=0.677	0.318 * p=0.017
Rating 40-49			1.504 p=0.280	0.396 p=0.060
Rating 50			1.465 p=0.229	0.432 p=0.059
Rating 51-59			1.364 p=0.427	0.373 * p=0.048
Rating 60-69			1.524 p=0.198	0.375 * p=0.030
Rating 70-79			1.737 p=0.126	0.464 p=0.105
Rating 80-89			3.042 ** p=0.003	0.767 p=0.583
Rating 90-95			3.136 ** p=0.006	0.825 p=0.710
Rating 96-100			3.071 ** p=0.006	0.799 p=0.662
Avg rating, start of call			1.010 * p=0.024	1.005 p=0.305
rho	0.021	0.010	0.018	0.013
(se)	(0.009)	(0.007)	(0.009)	(0.008)

* p<0.05, ** p<0.01, *** p<0.001

Appendix 4I. Boundaries for Each Tertile

Table I1. Minimum and Maximum Predicted Probabilities of the Low, Medium and High Tertiles for Each Model

				Classic			
Date model was run	Day of data collection	Model n	Cases	Minimum predicted prob. for last contact	Maximum for low tertile	Maximum for medium tertile	Maximum predicted prob. for last contact
November 5	7	551	389	0.002	0.297	0.524	0.794
November 12	14	1069	575	0.008	0.263	0.453	0.744
November 17	19	1232	602	0.004	0.221	0.433	0.718
November 20	21	1406	721	0.006	0.254	0.445	0.720
November 21	22	1508	787	0.005	0.267	0.447	0.709
November 28	28	2217	1104	0.006	0.244	0.399	0.675
November 29	29	2303	1134	0.006	0.241	0.394	0.670
December 3	33	2506	1179	0.007	0.223	0.385	0.656
December 7	37	2661	1206	0.009	0.214	0.376	0.652
December 8	38	2804	1242	0.005	0.187	0.368	0.655
December 12	42	3029	1288	0.006	0.184	0.355	0.647
December 13	43	2943	1261	0.006	0.196	0.355	0.649

	Ratings-only				Classic+			
Date model was run	Minimum predicted prob. for last contact	Maximum for low tertile	Maximum for medium tertile	Maximum predicted prob. for last contact	Minimum predicted prob. for last contact	Maximum for low tertile	Maximum for medium tertile	Maximum predicted prob. for last contact
November 5	0.081	0.308	0.458	0.615	0.003	0.264	0.545	0.831
November 12	0.074	0.273	0.428	0.564	0.014	0.260	0.463	0.786
November 17	0.043	0.246	0.396	0.563	0.005	0.224	0.434	0.768
November 20	0.056	0.259	0.407	0.574	0.007	0.251	0.449	0.796
November 21	0.057	0.264	0.407	0.568	0.006	0.263	0.440	0.774
November 28	0.050	0.255	0.362	0.547	0.007	0.238	0.395	0.735
November 29	0.045	0.254	0.359	0.542	0.007	0.232	0.387	0.727
December 3	0.041	0.252	0.342	0.542	0.008	0.217	0.379	0.725
December 7	0.031	0.241	0.333	0.529	0.009	0.214	0.374	0.714
December 8	0.029	0.232	0.323	0.531	0.005	0.188	0.366	0.718
December 12	0.024	0.222	0.319	0.533	0.006	0.180	0.355	0.727
December 13	0.028	0.229	0.320	0.534	0.007	0.192	0.360	0.730

Appendix 4J. Detailed Case Histories for a Selection of Cases Corresponding to Figure 4.9

Case 10 had two contacts, on November 19 and 20. The case provided an interview on the second contact. Although the case was included in the analysis and a predicted probability was generated on November 20, it is no longer active after that date. This case does not appear in figure 4.9 because the predicted probabilities are only shown for active cases that have not yet cooperated.

Cases 11, 12, 15, and 16 all interviewed on the first contact. They are not part of the analysis because the first contact was dropped for each case. They do not appear in figure 4.9.

Case 14 had three contacts, on November 24, 26, and 29, and cooperated on the last contact. Daily monitoring was conducted on November 21, 28 and 29. The first date that this case is part of the daily modeling is November 28, when there have been two contacts. Since case 14 did not cooperate on November 28, it is included in the predictions of cases that will cooperate at the next contact for November 28. Since case 14 cooperated on November 29, it is no longer active and only shows one predicted probability in figure 4.9.

Case 18 has three contacts and, although difficult to differentiate, two predicted probabilities shown in figure 4.9. This case was contacted twice on November 20 and provided an interview on November 22. Therefore, the case was an active case with predicted probabilities for the daily monitoring dates of November 20 and 21.

Case 48 has thirteen contacts between October 30 and November 24 but never cooperated. Figure 4.9 shows twelve predicted probabilities for the twelve dates in the daily monitoring (some of the probabilities are similar so the markers are not distinguishable). The first date of monitoring on November 5, uses the data from the second and third contacts on October 31 and November 2.

Appendix 4K. Success Rates for High, Medium, and Low Probability Tertiles

Table K1. Percent of High Probability Cases that Cooperated at the Next Contact for Each Model

	Classic		Ratings-only		Classic+	
Date model was run	**Number of active cases in high probability tertile**	**Percent that participated at next contact**	**Number of active cases in high probability tertile**	**Percent that participated at next contact**	**Number of active cases in high probability tertile**	**Percent that participated at next contact**
November 5	25	20%	31	23%	25	28%
November 12	28	11%	30	23%	23	13%
November 17	19	0%	29	14%	18	0%
November 20	43	12%	32	13%	36	11%
November 21	47	6%	34	9%	41	7%
November 28	54	9%	54	11%	49	6%
November 29	49	8%	56	18%	44	14%
December 3	48	8%	59	12%	42	10%
December 7	45	4%	61	13%	39	5%
December 8	39	0%	62	8%	32	0%
December 12	37	0%	54	4%	30	0%
December 13	37	0%	52	2%	31	0%

Table K2. Percent of Medium Probability Cases that Cooperated at the Next Contact for Each Model

	Classic		Ratings-only		Classic+	
Date model was run	**Number of active cases in medium probability tertile**	**Percent that participated at next contact**	**Number of active cases in medium probability tertile**	**Percent that participated at next contact**	**Number of active cases in medium probability tertile**	**Percent that participated at next contact**
November 5	51	33%	54	19%	49	24%
November 12	57	16%	70	3%	65	14%
November 17	50	10%	57	2%	51	6%
November 20	65	15%	89	15%	67	16%
November 21	77	17%	102	14%	79	14%
November 28	116	8%	133	8%	126	9%
November 29	123	10%	119	4%	131	8%
December 3	119	8%	107	3%	131	6%
December 7	109	6%	132	3%	128	9%
December 8	117	7%	146	6%	134	7%
December 12	120	5%	144	3%	128	5%
December 13	118	3%	152	3%	129	2%

Table K3. Percent of Low Probability Cases that Cooperated at the Next Contact for Each Model

	Classic		Ratings-only		Classic+	
Date model was run	**Number of active cases in low probability tertile**	**Percent that participated at next contact**	**Number of active cases in low probability tertile**	**Percent that participated at next contact**	**Number of active cases in low probability tertile**	**Percent that participated at next contact**
November 5	102	14%	93	20%	104	16%
November 12	131	7%	116	10%	128	7%
November 17	149	5%	132	5%	149	6%
November 20	165	4%	152	3%	170	4%
November 21	174	4%	162	4%	178	5%
November 28	254	6%	237	5%	249	6%
November 29	265	5%	262	6%	262	5%
December 3	281	4%	282	5%	275	5%
December 7	294	4%	255	4%	281	3%
December 8	317	5%	265	4%	307	5%
December 12	311	3%	270	3%	310	3%
December 13	333	3%	284	3%	328	3%

References

Achatz, Juliane, and Mark Trappmann. 2011. "Arbeitsmarktvermittelte Abgänge aus der Grundsicherung: der Einfluss von personen- und haushaltsgebundenen Barrieren." IAB-Discussion Paper 02/2011. Nürnberg.

American Association for Public Opinion Research. 2009. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 6th ed. AAPOR.

Bates, Nancy, James Dahlhamer, Polly Phipps, Adam Safir, and Lucilla Tan. 2010. "Assessing Contact History Paradata Quality across Several Federal Surveys." *JSM Proceedings of the Survey Research Methods Section of the American Statistical Association,* 91-105.

Beaumont, Jean-Francois. 2005. "On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment." *Survey Methodology,* 31: 227-31.

Beerten, Roeland, and Stephanie Freeth. 2004. "Exploring Survey Nonresponse in the UK: The Census-Survey Nonresponse Link Study." *Working Paper, Office for National Statistics, London*, 1–16.

Bethlehem, Jelke. 1988. "Reduction of Nonresponse Bias Through Regression Estimation." *Journal of Official Statistics,* 4: 251-60.

Bethmann, Arne, and Daniel Gebhardt, eds. 2011. *User Guide "Panel Study Labour Market and Social Security" (PASS), Wave 3.* FDZ Datenreport, 04/2011 EN, Nürnberg: Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research. http://www.iab.de/389/section.aspx/Publikation/k110824a05

Biemer, Paul P., and Andy Peytchev. 2012. "Census Geocoding for Nonresponse Bias Evaluation in Telephone Surveys." Public Opinion Quarterly 76: 432-52.

Biemer, Paul P., Patrick Chen, and Kevin Wang. 2013. "Using Level-of-Effort Paradata in Non-response Adjustments with Application to Field Surveys." *J.Royal Statistical Society A* 176: 147-68.

Biewen, Martin. 2006. "Who are the Chronic Poor? An Econometric Analysis of Chronic Poverty in Germany." In *Research on Economic Inequality: Dynamics of Inequality and Poverty*, edited by John Creedy and Guyonne Kalb, 31-62. Oxford: JAI Press.

Bundesministerium für Arbeit und Soziales. 2013. *Lebenslagen in Deutschland. Der 4*. Armuts- und Reichtumsbericht der Bundesregierung. Berlin.

Browne, William J. 2009. *MCMC Estimation in MLwiN v2.1*. Centre for Multilevel Modeling, University of Bristol.

Campanelli, Pamela, Patrick Sturgis, and Susan Purdon. 1997. “Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates.” Technical report, The Survey Methods Centre at SCPR, London.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd edition. New York: Chapman & Hall.

Casas-Cordero, Carolina. 2010. “Neighborhood Characteristics and Participation in Household Surveys.” Ph.D. thesis, University of Maryland. Available at http://hdl.handle.net/1903/11255.

Casas-Cordero, Carolina, Frauke Kreuter, Yueyan Wang, and Susan Babey. 2013. “Assessing the Measurement Error Properties of Interviewer Observations of Neighborhood Characteristics.” *Journal of the Royal Statistical Society A.*, 176: 227-49.

Couper, Mick P. 1998. “Measuring Survey Quality in a CASIC Environment.” *JSM Proceedings, Survey Research Methods Section of the American Statistical Association*, 41-9.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. “Changes in Telephone Survey Nonresponse over the Past Quarter Century.” *Public Opinion Quarterly*, 69: 87-98.

Davies, Paul S., and T. Lynn Fisher. 2009. “Measurement Issues Associated with Using Survey Data Matched with Administrative Data from the Social Security Administration.” *Social Security Bulletin*.69(2):1-12.

de Leeuw, Edith, and Wim de Heer. 2002. “Trends in Household Survey Nonresponse: A Longitudinal and International Comparison.” In *Survey Nonresponse*, eds. R. Groves, D. Dillman, J. Eltinge, and R. J. A. Little. New York: Wiley.

Department for Communities and Local Government. 2011. *“The English Indices of Deprivation 2010.”* London: Her Majesty’s Stationery Office. Available at www.communities.gov.uk/publications/corporate/statistics/indices2010.

Department for Communities and Local Government. 2012a. “Number and Type of English Local Authorities. London: Her Majesty’s Stationery Office.” Available at www.communities.gov.uk/localgovernment/local/.

Department for Communities and Local Government. 2012b. *“Thinking of Buying Your Council Flat?”* London: Her Majesty’s Stationery Office. Available at www.communities.gov.uk/publications/housing/buyingcouncilflats.

DiSogra, Charles, J. Michael Dennis, and Mansour Fahimi. 2010. “On the Quality of Ancillary Data Available for Address-Based Sampling.” Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 4174-83.

Dugmore, Keith. 2010. “Information Collected by Commercial Companies: What Might be of Value to Official Statistics? The Case of the UK Office for National Statistics”. Working Paper Series No. 151. Berlin: RatSWD.

Durrant, Gabriele B., and Fiona Steele. 2009. “Multilevel Modeling of Refusal and Noncontact in Household Surveys: Evidence from Six UK Government Surveys.” *Journal of the Royal Statistical Society, Series A* 172: 361–81.

Durrant, Gabriele B., Robert M. Groves, Laura Staetsky, and Fiona Steele. 2010. “Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys.” *Public Opinion Quarterly* 74:1–36.

Durrant, Gabriele B., Julia D’Arrigo, and Gerrit Müller. 2013. “Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys”. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter. New York: Wiley. p.281-308.

Eckman, Stephanie, Jennifer Sinibaldi, and Aleksa Möntmann-Hertz. 2013. “Can Interviewers Rate the Likelihood of Cases to Cooperate?” *Public Opinion Quarterly* 77: 561-73.

English, Ned, Ipek Bilgen, and Lee Fiorio. 2012. “Coverage Implications of Targeted Lists for Rare Populations.” Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 4521-28.

Fischhoff, Baruch, and Wändi Bruine de Bruin. 1999. “Fifty-Fifty = 50%?” *Journal of Behavioral Decision Making* 12: 149–63.

Frechette, Guillaume R. 2001. *Stata Technical Bulletin*. StataCorp LP, 10: 261-66.

Freeth, Stephanie, Catherine Kane, and Allison Cowie. 2002. “Survey Interviewer Attitudes and Demographic Profile: Preliminary Results from the 2001 ONS Interviewer Attitudes Survey.” *Working Paper, Office for National Statistics, London*, 1–18.

Fuchs, Benjamin. 2012. *Gründe für den Arbeitslosengeld-II-Bezug: Wege in die Grundsicherung.* IAB-Kurzbericht 25/2012. Nürnberg.

Fuller, Wayne A. 1987. *Measurement Error Models.* NewYork: Wiley.

Gebhardt, Daniel, Gerrit Müller, Arne Bethmann, Mark Trappmann, Bernhard Christoph, Christine Gayer, Bettina Müller, Anita Tisch, Bettina Siflinger, Hans Kiesl, Bernadette Huyer-May, Juliane Achatz, Claudia Wenzig, Helmut Rudolph, Tobias Graf, and Anika Biedermann. 2009. *Codebook and Documentation of the Panel Study “Labour Market and Social Security” (PASS). Volume I: Introduction and Overview, Wave 2 (2007/2008).* Nürnberg: Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research. http://fdz.iab.de/187/section.aspx/Publikation/k121214303

Groen, Jeffrey A. 2012. “Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures.” *Journal of Official Statistics* 28: 173-98.

Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly*, 70: 646-75.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys.* New York: Wiley.

Groves, Robert M., and Steven G. Heeringa. 2006. “Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs.” *Journal of the Royal Statistical Society Series A: Statistics in Society,* 169: 439 -57.

Groves, Robert M., and Emilia Peytcheva. 2008. “The Impact of Nonresponse Rates on Nonresponse Bias.” *Public Opinion Quarterly*, 72: 167- 89.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. “Leverage-Saliency Theory of Survey Participation: Description and an Illustration.” *Public Opinion Quarterly* 64: 299-308.

Hosmer, David W. Jr, and Stanley Lemeshow. 2000. *Assessing the fit of the model Applied Logistic Regression.* 2nd ed. Hoboken: John Wiley & Sons, Inc; pp.143–67.

Hubbard, Frost, and James Lepkowksi. 2009. “Experian Database Review.” Internal memorandum for the Survey Research Center Technical Infrastructure Group. July 16, 2009; Ann Arbor, MI.

Huynh, Minh, Kalman Rupp, and James Sears. 2002. *The Assessment of Survey of Income and Program Participation Benefit Data Using Longitudinal Administrative Records. Survey of Income and Program Participation Report No. 238.* Washington, DC: US Census Bureau.

Jowell, Roger, Caroline Roberts, Rory Fitzgerald, and Gillian Eva. 2007. *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*. London: Sage Publications.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. “Weighting Methods.” *Journal of Official Statistics,* 19: 81-97.

Kapteyn, Arie, and Jelmer Y. Ypma. 2007. “Measurement Error and Misclassification: A Comparison of Survey and Administrative Data.” *Journal of Labor Economics* 25: 513-51.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey.” *Public Opinion Quarterly*, 70: 759-79.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. “Consequences of Reducing Nonresponse in a National Telephone Survey.” *Public Opinion Quarterly*, 64: 125-48.

Kirgis, Nicole G., and James M. Lepkowski. 2013. “Designs and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006-2010 National Survey of

Family Growth." In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter. 123-44. Hoboken: Wiley.

Kreuter, Frauke. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter. Hoboken: Wiley.

Kreuter, Frauke, and Carolina Casas-Cordero. 2010. "Paradata." Working Paper No. 136, German Council for Social and Economic Data (RatSWD).

Kreuter, Frauke, and Kristen Olson. 2011. "Multiple Auxiliary Variables in Nonresponse Adjustment." *Sociological Methods and Research,* 40: 311-22.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010a. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74: 880-906.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trina M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivellore E. Raghunathan. 2010b. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys." *J.Royal Statistical Society A,* 173: 389-407.

Laflamme, Francois, Mike Maydan, and Andrew Miller. 2008. "Using Paradata to Actively Manage Data Collection Survey Process", *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, Denver, Colorado, USA.

Lepkowski, James M., William D. Mosher, Karen E. Davis, Robert M. Groves, and John Van Hoewyk. 2010. "The 2006–2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey." National Center for Health Statistics, Vital and Health Statistics 2(150).

Little, Roderick J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54, 139 – 57.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data.* New York: Wiley.

Little, Roderick J. A., and Sonja Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology,* 31: 161-68.

McCulloch, Susan, Frauke Kreuter, and Stephanie Calvano. 2010. "Interviewer Observed vs. Reported Respondent Gender: Implications on Measurement Error." *Paper presented at the 2010 annual meeting of the American Association for Public Opinion Research*, Chicago, IL.

McFall, Stephanie L. 2011. *Understanding Society—The UK Household Longitudinal Study, Wave 1, 2009–2010, User Manual*. Colchester, UK: University of Essex.

microm Consumer Marketing. 2013. microm Datenhandbuch. Nuess, Germany: microm Micromarketing-Systeme und Consult GmbH.

Miller, Peter. 2013. “Goals and General Framework.” *Presentation made at the workshop for Advances in Adaptive and Responsive Survey Design, Heerlen, Netherlands, December 9, 2013.*

Oemmelen, Guido. 2012. “Microgeographische Daten der microm. Entwicklung, Nutzen und Relevanz sowie Beispiele aus der sozialwissenschaftlichen Forschung.” Paper presented at Forschungskolloquiums am Bundesinstitut für Bevölkerungsforschung (BiB), Wiesbaden, Germany: February 8. http://www.bib-demografie.de/SharedDocs/Publikationen/DE/Veranstaltungen/Forschungskolloquium/Praesentationen/2012/p_2012_02_08_oemmelen.pdf

Office for National Statistics, General Register Office for Scotland, and Northern Ireland Statistics and Research Agency. 2004. *Census 2001 Definitions*. London: Her Majesty’s Stationery Office. Available at www.statistics.gov.uk.

O’Hare, Barbara C. 2012. “The Use of Paradata to Improve Survey Quality: Census Bureau.” *Paper presented at the 2012 Federal Computer Assisted Survey Information Collection Workshop, Washington, DC*. Available at https://fedcasic.dsd.census.gov/fc2012/.

Olson, Kristen, and Robert M. Groves. 2012. “An Examination of Within-Person Variation in Response Propensity over the Data Collection Field Period.” *Journal of Official Statistics*, 28: 29-51.

Olson, Kristen, Jolene D. Smyth, and Heather M. Wood. 2012. “Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination.” *Public Opinion Quarterly* 76: 611 – 35.

O’Muircheartaigh, Colm, and Pamela Campanelli. 1999. “A Multilevel Exploration of the Role of Interviewers in Survey Non-response.” *J.Royal Statistical Society A* 162: 437-46.

Pencina, Michael J., Ralph B. D’Agostino Sr, Ralph B. D’Agostino Jr, and Ramachandran S. Vasan. 2008. “Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond.” *Statistics in Medicine* 27: 157–72.

Peytchev, Andy, Rodney K. Baxter, and Lisa R. Carley-Baxter. 2009. “Not All Survey Effort is Equal: Reduction of Nonresponse Bias and Nonresponse Error.” *Public Opinion Quarterly*, 73: 785-806.

Peytcheva, Emilia, and Robert M. Groves. 2009. “Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates.” *Journal of Official Statistics* 25: 193-201.

Pickering, Kevin, Roger Thomas, and Peter Lynn. 2003. “*Testing the Shadow Sample Approach for the English House Condition Survey*.” Report to the Office of the Deputy Prime Minister. London: National Centre for Social Research.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. 3rd ed. Vol. 2. College Station, TX: Stata Press.

Rasbash, Jon, Fiona Steele, William J. Browne, and Harvey Goldstein. 2009. *A User's Guide to MLwiN v2.1*. Centre for Multilevel Modelling, University of Bristol.

Rastogi, Sonya, and Amy O'Hara. 2012. *2010 Census Match Study.* Washington DC: Center for Administrative Records Research and Applications, U.S. Census Bureau.

Riphahn, Regina T., Monika Sander, and Christoph Wunder. 2013. "The Welfare Use of Immigrants and Natives in Germany: the Case of Turkish Immigrants". *International Journal of Manpower* 34: 70-82.

Sakshaug, Joseph W., and Frauke Kreuter. 2012. "Assessing the Magnitude of Non-Consent Bias in Linked Survey and Administrative Data." *Survey Research Methods* 6: 113-22.

Schnell, Rainer. 2012. *Survey-Interviews. Methoden Standardisierter Befragungen.* Wiesbaden: VS-Verlag.

Schnell, Rainer, and Frauke Kreuter. 2005. "Separating Interviewer and Sampling-Point Effects." *Journal of Official Statistics* 21: 389–410.

Schouten, Barry. 2013. "Workshop General Introduction. "Presentation at Workshop Advances in Adaptive and Responsive Survey Design, Heerlen, Netherlands.

Schouten, Barry, Melania Calinescu, and Annemieke Luiten. 2013. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology* 39: 29-58.

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. 2009. "Indicators for the Representativeness of Response." *Survey Methodology* 35: 01–13.

Schräpler, Jörg-Peter, Jürgen Schupp, and Gert G. Wagner. 2010. "Individual and Neighborhood Determinants of Survey Nonresponse – An Analysis Based on a New Subsample of the German Socio-Economic Panel (SOEP), Microgeographic Characteristics and Survey-Based Interviewer Characteristics." SOEP papers No. 288. Berlin: German Institute for Economic Research (DIW).
http://www.diw.de/documents/publikationen/73/diw_01.c.354686.de/diw_sp0288.pdf

Singer, Judith D., and John B. Willett. 1993. "It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events." *Journal of Educational and Behavioral Statistics* 18: 155-95.

Sinibaldi, Jennifer. 2010. "Measurement Error in Objective and Subjective Interviewer Observations." *Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Chicago*.

Sinibaldi, Jennifer, Gabriele B. Durrant, and Frauke Kreuter. 2013. "Evaluating the Measurement Error of Interviewer Observed Paradata." *Public Opinion Quarterly*, 77: 173-93.

Sinibaldi, Jennifer, Mark Trappmann, and Frauke Kreuter. *forthcoming*. “Which is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations?” *Public Opinion Quarterly*.

Spiegelhalter, David J., Nicky G. Best, Bradley P. Carlin, and Angelika van der Linde. 2002. “Bayesian Measures of Model Complexity and Fit (with Discussion).” *Journal of the Royal Statistical Society. Series B*. 64, 4: 583-640.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Trappmann, Mark. 2011. “Weighting. In User Guide ‘Panel Study Labour Market and Social Security’ (PASS), Wave 3.” In FDZ Datenreport, 04/2011 EN, edited by Arne Bethmann and Daniel Gebhardt, 51-61. Nürnberg.

Trappmann, Mark, Stefanie Gundert, Claudia Wenzig, and Daniel Gebhardt. 2010. “PASS: A Household Panel Survey for Research on Unemployment and Poverty.” *Schmollers Jahrbuch: Journal of Applied Social Science Studies*. 130: 609-22

Tutz, Gerhard, and Jan Gertheiss. 2013. “Rating Scales as Predictors: The Old Question of Scale Level and Some Answers”. *Psychometrika,* 1-20.

Tutz, Gerhard, and Margret-Ruth Oelker. 2014. “Modeling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures”. Technical report 156. Munich: University of Munich, Department of Statistics.

Tversky, Amos and Daniel Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185(4157): 1124-31.

Wagner, James. 2013. “Adaptive Contact Strategies in Telephone and Face-to-Face Surveys.” *Survey Research Methods*, 7: 45-55.

Wagner, James, and Frost Hubbard. *forthcoming*. “Producing Consistent Estimates of Propensity Models during Data Collection.” *Journal of Survey Statistics and Methodology*.

Wagner, James, Brady T. West, Nicole Kirgis, James M. Lepkowski, William G. Axinn, and Shonda Kruger Ndiaye. 2012. “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection.” *Journal of Official Statistics* 28: 477-99.

Ware, James H. 2006. “The Limitations of Risk Factors as Prognostic Tools.” *New England Journal of Medicine* 355: 2615-7.

Weich, Scott, Elizabeth Burton, Martin Blanchard, Martin Prince, Kerry Sproston, and Bob Erens. 2001. “Measuring the Built Environment: Validity of a Site Survey Instrument for Use in Urban Settings.” *Health and Place* 7: 283–92.

West, Brady T. 2010. “A Practical Technique for Improving the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth.” NSFG Survey

Methodology Working Papers, No. 10–013. Ann Arbor: University of Michigan, Institute for Social Research.

West, Brady T. 2013a. “An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth.” *Journal of the Royal Statistical Society: Series A*, 176: 211-25.

West, Brady T. 2013b. “The Effects of Errors in Paradata on Weighting Class Adjustments: A Simulation Study”. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter, 361-88. New York: Wiley.

West, Brady T., and Robert M. Groves. 2013. “A Propensity-Adjusted Interviewer Performance Indicator.” *Public Opinion Quarterly* 77: 352-74.

West, Brady T., and Roderick J. A. Little. 2013. “Non-response Adjustment of Survey Estimates Based on Auxiliary Variables Subject to Error.” *Applied Statistics,* 62(2): 1-19.

West. Brady T., and Kristen Olson. 2010. “How Much of Interviewer Variance is Really Nonresponse Error Variance?” *Public Opinion Quarterly* 74: 1004-26.

West, Brady T., and Jennifer Sinibaldi. 2013. “The Quality of Paradata: A Literature Overview.” In *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by Frauke Kreuter, 339-60. Hoboken: Wiley.

West, Brady T., Frauke Kreuter, and Ursula Jaenichen. 2013. “‘Interviewer’ Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?” *Journal of Official Statistics,* 29: 277–97.

West, Brady T., Frauke Kreuter, and Mark Trappmann. 2012. “Observational Strategies Associated with Increased Accuracy of Interviewer Observations in Employment Research.” Paper presented at the annual meeting of the American Association of Public Opinion Research, Orlando, Florida.

West, Brady T., Frauke Kreuter, and Mark Trappmann. *forthcoming*. “Is the Collection of Interviewer Observations Worthwhile in an Economic Panel Survey? New Evidence from the German Labor Market and Social Security (PASS) Study.” *Journal of Survey Statistics and Methodology.*

Wood, Simon N. 2006.*Generalized Additive Models: An Introduction with R.* London: Chapman & Hall.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Sinibaldi, Jennifer

Name, Vorname

Ort, Datum

Unterschrift Doktorand/in

Formular 3.2