Chapter I

Introduction

Declining survey response rates continue to plague survey research organizations worldwide (Baruch and Holtom, 2008; Biener et al., 2004; Cull et al., 2005; Curtin et al., 2005; de Leeuw and de Heer, 2002; Tolonen et al., 2006). Fortunately, recent methodological work (Groves, 2006; Groves and Peytcheva, 2008; Keeter et al., 2000) has found that decreases in survey response rates do not necessarily lead to increases in the nonresponse bias of survey estimates. Instead, the bias in survey estimates introduced by nonresponse arises from correlations between the response propensities of individuals in a population and the survey variables of interest (Bethlehem, 2002). The most effective auxiliary variables for reducing the bias in survey estimates that arises from nonresponse therefore need to have three important properties: 1) they are available for both respondents and nonrespondents; 2) they are correlated with survey response indicators; and 3) they are correlated with survey variables of interest (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Lessler and Kalsbeek, 1992; Little and Vartivarian, 2005). Unfortunately, auxiliary variables available to survey researchers for both respondents and nonrespondents seldom satisfy the third condition (i.e., they have weak correlations with the survey variables).

Because of this problem, surveys conducting in-person household interviews often request that field interviewers observe and record selected characteristics (i.e., characteristics correlated with key survey variables) of all sampled households. Kreuter et al. (2010) present three examples: the American National Election Study (ANES), the Continuous National Survey of Family Growth (NSFG), and the European Social Survey (ESS). The recently completed seventh cycle of the NSFG, for instance, asked interviewers to judge whether households had children under the age of 15 present, based on observations of sampled housing units made prior to screening interviews. In addition, NSFG interviewers were asked to judge (on the doorstep) whether potential respondents selected from completed screening interviews were currently sexually active (Groves et al., 2009).

The third chapter of this dissertation will show that these two judgments are correlated with both indicators of response to the main NSFG interview and multiple key NSFG variables, which makes these interviewer judgments attractive candidates for making post-survey nonresponse adjustments. However, one could argue that interviewer observations like these and other *paradata*, or data collection process variables (Beaumont, 2005; Couper, 1998; Couper and Lyberg, 2005), are really *random* variables (in that their values could vary across replications of the survey process using the same sample). Fortunately, Beaumont (2005) has shown that using these kinds of "random" variables in post-survey nonresponse adjustments results in no additional bias or components of variance in the resulting estimators (meaning that they can be used in the same way as other fixed auxiliary variables in the adjustments).

Unfortunately, the practice of using interviewer observations to make nonresponse adjustments has outpaced the theory underlying this methodology. Auxiliary variables used for making nonresponse adjustments in practice are typically characteristics of area sampling units (Lessler and Kalsbeek, 1992), which may not be related to survey variables and / or response indicators but are generally measured without error. Interviewer observations, on the other hand, generally rely on judgment and estimation and will thus be prone to error[1]. For instance, Moon (1999) describes the practice of having exit poll interviewers guess the political candidates for whom all attempted individuals voted, which is a procedure that (understandably) can result in an auxiliary variable that is fraught with error.

The existing literature on interviewer observations has not yet fully considered the error properties of these auxiliary variables and the implications of these error properties for the effectiveness of nonresponse adjustments based (in part) on the observations. More generally, the implications of errors in *any* auxiliary variables for the effectiveness of nonresponse adjustments have received very little research attention. Even more broadly, substantial errors in auxiliary variables have the potential to impact a wide variety of survey methodologies aside from the development of nonresponse adjustments. Very few studies have addressed these potential impacts, making this a ripe area for research. As pointed out in a recent international report on the use of auxiliary data to adjust for

[1] Funder (1987) argues that these types of errors, based on judgments made in a real-world social environment, should be referred to as "mistakes," or "misjudgments of more poorly defined, real-life stimuli." This research will continue to refer to these "mistakes" as errors, given that the interviewers are attempting to collect information that sampled persons could provide (in theory) if responding to the survey request.

nonresponse bias in surveys, "Research is also needed on the verification and standardization of observational paradata and how to maximize their reliability" (Smith, 2011, p. 393).

As an initial step in addressing the methodological concerns that arise from working with error-prone auxiliary variables, this dissertation reviews the existing literature on the quality and utility of interviewer observations, presents results from three research studies designed to address these important gaps in the literature, discusses the implications of the study results for survey practice, and concludes with directions for future research in this area. The University of Michigan Institutional Review Board (IRB) approved all studies presented in this dissertation.

Chapter II

Review of the Literature

This chapter presents a comprehensive review of the existing survey methodological and sociological literature examining 1) the effectiveness of nonresponse adjustments based on interviewer observations; 2) the error properties of interviewer observations; 3) possible predictors of the accuracy in interviewer observations; and 4) the implications of errors in interviewer observations for the effectiveness of nonresponse adjustments.

The Effectiveness of Nonresponse Adjustments Based on Interviewer Observations

Few studies have formally evaluated the effectiveness of nonresponse adjustments to survey estimates based on interviewer observations. Dillard and Ford (1984) described a procedure tested by the U.S. Department of Agriculture (USDA) for imputing farm data on the presence of hogs based on enumerator observations, and demonstrated that this procedure did a better job of repairing nonresponse errors than other operational procedures in use at the time. Groves and Heeringa (2006) used interviewer observations in the National Survey of Family Growth (NSFG) to identify and prioritize active cases with a high probability of response in a two-phase design (where the second phase featured an intensive follow-up of active nonrespondents), ultimately increasing response rates while reducing follow-up costs. Whether or not the increase in response rates lead to

a decrease in nonresponse bias, however, remains an open question, given existing literature suggesting that response rates are not correlated with the nonresponse bias of survey estimates (Groves, 2006; Groves and Peytcheva, 2008). Blom (2009) reported strong associations of interviewer observations on features of residential buildings with cooperation in the European Social Survey (ESS), although it was not clear how large of a role the observations played in reducing nonresponse bias relative to other ESS contact data. Analyzing data from the 2007 British Gambling Prevalence Survey (BGPS), Scholes et al. (2008) reported that interviewer observations on external household features (e.g., existence of physical barriers to entry) were predictive of both household response and key survey variables of interest (e.g., frequent gambling), but found that reductions in nonresponse bias were not consistent across different variables when using these observations to make nonresponse adjustments.

Kreuter et al. (2010) studied the effectiveness of nonresponse adjustments based on interviewer observations in the NSFG, the American National Election Studies (ANES), and the ESS. These authors found that the interviewer observations generally tended to have weak correlations with response propensity, and that nonresponse adjustments only tended to shift estimates when the interviewer observations had stronger correlations with the survey variables. Gonzalez and Kreuter (2010) analyzed data from the 2009 Consumer Expenditure (CE) quarterly interview survey, and found that interviewer observations of housing tenure and principal components reflecting interviewer-perceived survey participation concerns of contacted respondents (collected as part of the U.S. Census Bureau's Contact History Instrument, or CHI) had weak negative correlations

with response propensity and only small correlations with several key variables (with the exception of a few moderate correlations of housing tenure with key variables). Consistent with results presented by Kreuter et al. (2010), these authors found that larger absolute changes in the correlation of predicted response propensities with key survey variables engendered by the addition of interviewer observations to nonresponse adjustments were associated with greater shifts in estimates of means for key variables.

Other existing studies have simply examined associations of interviewer observations with response propensity and / or key survey variables, without studying the effectiveness of nonresponse adjustments to estimates based on the observations. Blom et al. (2011) report that selected interviewer observations of housing and neighborhood conditions in the ESS are strong predictors of both successful contact and successful cooperation conditional on contact. Durrant et al. (2011a) and D'Arrigo et al. (2011) fitted multilevel multinomial logistic regression models to call outcomes from six major government surveys in the United Kingdom, and found that a variety of time-varying (e.g., method of contact) and time-invariant (i.e., neighborhood and household) interviewer observations were predictive of cooperation in the six surveys, suggesting that the observations could be useful for informing responsive survey designs. In a related paper, Durrant et al. (2011b) showed that interviewer observations are predictive of establishing contact with a household, and can be used to predict the best times to contact a household (even when taking other household-level information into account). Durrant et al. (2010, p. 13) found significant associations of interviewer observations of housing unit conditions with response propensity in a multivariate model, and this finding has also been reported by

Durrant and Steele (2009) and Lynn (2003). Durrant and Steele (2009) also reported associations of neighborhood safety observations with response propensity.

Casas-Cordero (2010b) found that 1) interviewer observations of the physical environment of neighborhoods in Los Angeles County were correlated with response propensity; 2) these observations mediated relationships of neighborhood SES with cooperation; and 3) the physical environment observations were moderate-to-strong correlates of multiple key survey variables. Campanelli et al. (1997, Section 4.6) found significant associations of several interviewer observations (including approximate size of household, presence of a security device, and being occupied with children) with response propensity in a multivariate model. Lynn (2003) found several significant associations of interviewer observations with response propensity, and reported that the observations were easy for interviewers to collect. Maitland et al. (2009) found that interviewer observations collected on the CHI had stronger correlations with response propensity than with variables of interest in the National Health Interview Survey (NHIS), and that an interviewer observation on a health-related measure had the strongest associations with key NHIS variables.

Collectively, this relatively small literature therefore suggests that correlations of interviewer observations with response propensity tend to vary for different surveys, and that the most effective interviewer observations for nonresponse adjustment are on correlates of key variables, as suggested by existing theory (Little and Vartivarian, 2005). The first study in this dissertation (Chapter 3) examines the effectiveness of nonresponse

adjustments using interviewer observations in the NSFG in more detail. In general, future research needs to focus on methods for quantifying the reductions in the nonresponse bias of key estimates when using interviewer observations for nonresponse adjustments.

The Error Properties of Interviewer Observations

Given that interviewer observations are typically judgments and estimates made by interviewers, frequent errors in the observations could be one reason for the limited effectiveness of nonresponse adjustments based on the observations reported in the literature to date. However, the existing literature has only begun to study the error properties of interviewer observations.

Initial validation studies have suggested that the prevalence of errors in these observations is by no means negligible. Preliminary analyses of NSFG data (Groves et al., 2007), for example, have found that doorstep interviewer judgments of whether screened respondents are currently sexually active are only 75-80% consistent with eventual respondent reports of sexual activity, and that false positive observations (i.e., judging that a person is currently sexually active when they report that they are not) are much more likely than false negative observations. Testing the potential effectiveness of the *shadow sample* approach for the English House Condition Survey (EHCS), Pickering and colleagues (2003) found accuracy rates for interviewer observations of various housing unit types ranging from 46% (for privately rented dwellings; i.e., 46% of privately rented dwellings based on actual household reports were observed to be privately rented dwellings) to 89% (for owner-occupied dwellings). In addition,

interviewer observations regarding the age of the shadow housing units ranged from 62% accurate (built 1945-1964) to 81% accurate (built before 1919), and observations for housing units defined by both type and age ranged from 7% accurate (registered social landlord dwellings built between 1919 and 1944) to 100% accurate (local authority flats built before 1919). This work, however, did not consider the types of error (i.e., false positives or false negatives) being made in the observations, and suggested that simulations should be used to examine the implications of the errors for sampling purposes.

A few existing validation studies have examined the frequencies of error in interviewer observations of respondent ethnicity. Hahn et al. (1996) analyzed data from the first National Health and Nutrition Examination Survey (NHANES I) and a follow-up data collection (the NHANES I epidemiologic follow-up study, or NHEFS). In the NHANES I, interviewers were asked to estimate the racial classification of NHANES I respondents (White, Negro, or Other), and the same respondents were asked to describe their ethnic background in the follow-up survey. Interviewers did a particularly inconsistent job of classifying Native Americans, with 61% classified as White and 38% classified as Negro. Nearly 99% of those identifying as African or Black were classified as Negro by the interviewers, while 91% of those identifying as European were classified as White by the interviewers. Similar findings were reported by Drury et al. (1980) in the NHIS, with Native Americans and Asians consistently receiving different classifications by interviewers tasked with estimating ethnicity. In this study, 25% of self-identified Asians were classified as White by Interviewers, while 74% of self-identified Asians were

classified as “Other.” In addition, 84% of American Indians were classified as White or Black, while 16% were classified as “Other.” These studies suggest that using interviewer observations of race for methodological purposes may be a poor choice, especially for studies including minority ethnic groups.

An initial validation study considering the prevalence of errors in interviewer judgments of the presence of children and the presence of smokers in households from the Health Survey for England (HSE) yielded accuracy rates of about 75% (Tipping and Sinibaldi, 2010), consistent with the findings of Groves and colleagues (2007). In the ESS, the accuracy of interviewer observations collected on contact forms has been shown to vary across both variables and countries (Stoop et al., 2010, Chapter 8). Analyzing tape recordings of the doorstep interactions of interviewers with potential respondents, Campanelli et al. (1997, Chapter 4) found very low agreement (less than 50%) of information in the recordings with contact observation data entered by the interviewers following the interactions, attributing this to possible interviewer memory errors. Another recent study evaluating the ability of trained interviewers to guess the gender of respondents on the telephone found only 92% accuracy (McCulloch et al., 2010). Additionally, Biemer et al. (2010) studied errors in the number of calls reported by interviewers for a sample unit, and found variance among interviewers in terms of what constituted an official “call” and how many calls were not being reported by interviewers. The existing validation studies in this area therefore suggest that there is wide variance in the accuracy of interviewer observations, largely depending on the feature being observed, and that some observations can in fact have very poor quality.

No validation studies of the errors in interviewer observations to date have examined trends in the accuracy of the observations over the life of a data collection, to determine whether interviewers improve their accuracy with more experience. Olson and Peytchev (2007) found that interviewer behaviors (i.e., length of survey administration) and perceptions do in fact change over the course of a data collection, but whether or not judgment accuracy improves as well remains an open question. Chapter 4 presents initial answers to this open question, considering results from multilevel models looking at the relationship of time since onset of data collection with the accuracy of interviewer judgments in the NSFG.

Other existing work in this area has examined *indirect* indicators of errors in the interviewer observations. Examples of these indirect indicators include consistency in both objective and subjective interviewer observations over time (Sinibaldi, 2010), inter-rater reliability of interviewer judgments, suggested by Funder (1987) as a useful tool for evaluating judgment accuracy (Alwin, 2008, p. 151; Casas-Cordero, 2010a; Casas-Cordero and Kreuter, 2008; Eckman, 2011; Kennickell et al., 2011, p. 6; Mosteller, 1944), within-area correlation in the observations (Casas-Cordero, 2010a), interviewer problems with collecting the observations (Pickering et al., 2003), substantial interviewer variance in subjective judgments given interpenetrated assignments (Feldman, 1951, p. 743), variance in perceptions of respondent skin color depending on interviewer race (Hill, 2002), and missing data rates for the observations (Kreuter et al., 2007; Lynn, 2003). For example, Alwin (2008, p. 151) reports results from a panel survey indicating

that inter-interviewer reliability of reports on factual household information tends to be high (0.85). This is largely consistent with Mosteller's (1944) work, with the exception of low inter-interviewer agreement on ratings of economic status (54% agreement). However, Alwin (2008) also reports that interviewer beliefs about characteristics of *respondents* tend to have much lower reliability. These studies using indirect indicators of error have also suggested that the quality of interviewer observations can be questionable, and the extant work in this area has consistently called for more direct validation studies of the errors in these observations (e.g., Kreuter et al., 2010; Kreuter et al., 2007; Yan and Raghunathan, 2007). The first study in this dissertation (Chapter 3) uses validation data to examine the error properties of interviewer observations collected in the NSFG in more detail.

Predictors of Judgment Accuracy from Social Psychology

While only a small number of existing studies have examined the accuracy and reliability of interviewer observations, no studies of surveys using in-person interviewing have considered *predictors* of accuracy in the interviewer observations, which could inform methods for *reducing* the error in these observations. Theories from the social psychology literature may provide insights to help guide survey methodologists in this pursuit, especially when interviewers are tasked with recording observations that are more subjective in nature. Considering doorstep judgments about characteristics of individuals (e.g., current sexual activity in the NSFG), interviewers tasked with collecting or recording these types of observations upon contact with sampled individuals are making what are known as "thin-slice" (or brief) observations of behaviors, based on first

impressions and intuitions (e.g., Ambady and Gray, 2002; Ambady et al., 1999; Winerman, 2005). Past studies have shown that the accuracy of these "thin-slice" judgments can be quite high (Ambady and Rosenthal, 1992; Winerman, 2005), but studies looking at *predictors* of the accuracy tend to be more rare.

Examining thin-slice judgments of sexual orientation, Ambady et al. (1999) showed that accuracy was a function of how dynamic nonverbal behaviors were (i.e., judgments based on silent videos had higher accuracy than those based on still photographs) and of features of the judges relevant to the judgments (gay men and lesbians were better at judging still photos and shorter videos than heterosexuals). Women have also been shown to judge sexual orientation more accurately than men (Berger et al., 1987). Other research in this area has shown that sadness has a consistent negative impact on the accuracy of a variety of thin-slice judgments (Ambady and Gray, 2002), which could have important practical implications for interviewers (i.e., interviewers should generally not collect observations when feeling sad or depressed). In addition, past work has shown that in cognitively demanding situations (such as listing and survey interviewing), judgments based on first impressions tend to have *higher* accuracy (Patterson and Stockbridge, 1998), which provides support for the idea of asking interviewers to make quick judgments based on first impressions.

Research in this area has also shown the importance of focusing on explicit, obvious, nonverbal visual cues (e.g., hairstyles for homosexual males) for increasing the confidence and accuracy in intuitive judgments of ambiguous social categories (e.g.,

sexual orientation), in addition to the correlation of judgment accuracy and judgment confidence (Patterson et al., 2001; Rule et al., 2008). Attending to multiple non-obvious (and more subtle) visual cues may actually lead to a reduced correlation of confidence with accuracy. These studies of the accuracy of subjective "thin-slice" judgments therefore lend support for the following hypothesis: Interviewers who focus on obvious, nonverbal visual cues (i.e., observable correlates of the feature being judged) when making specific judgments and also have characteristics relevant to the topic of the survey will have improved judgment accuracy.

Interviewer judgments of behavioral traits also fit into another psychological framework for social judgment known as the "zero-acquaintance paradigm" (Albright et al., 1988; Passini and Norman, 1966), where persons are asked to make judgments about the characteristics of complete strangers. Studies of zero-acquaintance situations have shown that people can in fact judge personality traits of strangers accurately based on brief observations (Ambady et al., 1999). Importantly, Ambady et al. (1995) studied *predictors* of judgment accuracy in zero-acquaintance situations, and found that judgment accuracy was higher in judges who were *female* and *less sociable*, and that male judges scoring higher on tests of nonverbal sensitivity and female judges scoring higher on tests of interpersonal perception also had higher accuracy. These findings suggest that ideal interviewers for making these types of behavioral judgments on complete strangers should be female, less sociable and high-performing on the aforementioned tests of perceptive ability. More survey methodological studies are certainly needed to test this hypothesis further.

Tversky and Kahneman (1974) also provide a theoretical framework describing mechanisms that lead to errors in judgments. Particularly relevant to the problem of errors in interviewer observations is what these authors refer to as the "representativeness" heuristic, where the probability that a judge estimates that someone falls into a certain class is based on the degree to which the person is representative of the stereotype for that class. Citing several experimental studies, these authors identify five problems with this heuristic that can lead to errors in judgments: 1) insensitivity to prior probabilities of outcomes (e.g., NSFG interviewers ignoring known proportions of the population that are sexually active, or percentages of households that have kids); 2) insensitivity to sample size (i.e., larger samples will be less likely deviate from known population features); 3) misconceptions of chance (i.e., interviewers should not simply guess that a person is sexually active just because the previous *n* persons were all judged to be sexually inactive); 4) insensitivity to predictability and illusions of validity (i.e., interviewers should make judgments based on reliable, independent, and relevant predictors of what is being judged); and 5) misconceptions of regression to the mean. This framework provides useful guidelines that could inform interviewer training and subsequent practice in the field, if observations are required.

Theoretical models for predicting the accuracy of personality and behavioral judgments as a function of judge (interviewer) and individual (respondent) characteristics have been proposed and applied in the psychological literature (Funder, 1995), but these approaches have not yet been applied by survey methodologists. In general, there has been minimal

prior work in the survey methodology literature examining multiple predictors of the accuracy in interviewer judgments (see McCulloch et al., 2010, for an example from a telephone survey), and exploratory work in this area is needed for generating research hypotheses. For example, McCulloch and colleagues (2010) found that more interviewer experience resulted in *higher* error rates when guessing respondent gender on the telephone, possibly indicating that interviewers with more experience might not take the observational task as seriously as newer interviewers with less experience. These authors also found high intra-interviewer correlations in judgment errors and significant interactions between the demographic features of respondents and interviewers when predicting accuracy. The second study in this dissertation (Chapter 4) aims to fill this gap in the survey methodological literature, considering a large number of respondent- and interviewer-level predictors of interviewer judgment accuracy in a personal interview survey (the NSFG).

While much of the social psychological literature in this area has focused on judge-level correlates of the accuracy of judgments regarding personality or behavioral traits, interviewers are often requested to observe household-level (rather than personal) features (e.g., Pickering et al., 2003). Current research on the visual cognition phenomenon known as *inattentional blindness*, or the failure to notice objects in plain sight when focused on a demanding visual task, has suggested that the difficulty of the observational task, rather than differences in individual ability, predicts failure to notice other objects (Simons and Jensen, 2009). This theory has been supported empirically by Casas-Cordero (2010a), who found that the probability of interviewers noting

neighborhood disorders was a function of selected neighborhood characteristics and not interviewer characteristics. In addition, Pickering et al. (2003) performed simple descriptive analyses examining variance in the accuracy of interviewer observations on housing unit types by urban/rural status, and found that the accuracy of the observations tended to be higher in rural areas. Interestingly, Kennickell (2003, p. 2123) reported that interviewers working on an area probability sample for the Survey of Consumer Finances (SCF) generally dedicated more effort to following cases in apartment buildings and largely Hispanic areas, while dedicating less effort to areas with higher incomes and higher proportions of people age 65 and over. This variance in levels of effort on the part of the interviewers, depending on the features of areas or housing units, could lead to variance in the quality of the observations across areas as well.

The existing visual cognition theory and initial empirical work in this area therefore provide support for a hypothesis that accuracy may suffer in areas where observations are more difficult, as opposed to being a function of interviewer ability. For instance, housing unit access problems and urban areas are hypothesized to have a negative impact on the accuracy of housing unit observations on the presence of young children, as interviewers may have a difficult time picking up important visual cues (e.g., children's toys) without having more ready access to the housing units. The second study in this dissertation (Chapter 4) presents analyses of a more general set of respondent- and interviewer-level predictors of accuracy in two NSFG interviewer judgments, identifying key correlates of error in both a behavioral judgment (current sexual activity) and a household-level judgment (presence of children under the age of 15).

Importantly, the design of the NSFG (and many other area probability samples) does not allow for interpenetrated assignment of subsamples of the full NSFG sample to interviewers (interviewers are typically assigned to work in a single PSU only for cost efficiency). Judgment accuracy may therefore *implicitly* vary across interviewers before data collection even begins, given that judgments may be more or less difficult depending on the features of a particular PSU. The inclusion of several PSU-level and neighborhood-level features as predictors in the models in Chapter 4 (e.g., variety of PSUs worked by an interviewer, neighborhood safety concerns, urban / rural segment, etc.) represents an attempt to eliminate any sources of variance in accuracy either between or within interviewers due to factors beyond their control, given the design of the NSFG. Remaining components of variance due to interviewers after taking these factors into account provide at best an approximation of the variance in accuracy introduced by the interviewer, but more research into methodologies for making inferences about interviewer variance in the case of non-interpenetrated sample assignments is needed.

In addition, past studies on person perception have assumed that although judges are aware of their ability to pick up nonverbal cues in making accurate judgments, they are unable to articulate the cues used (Smith et al., 1991). Other work in this area has indicated that asking judges to rate the confidence of their judgments (e.g., Pickering et al., 2003, p. 17) may impair cognitive mechanisms leading to more accurate judgments (Patterson et al., 2001). In 2010, the NSFG started collecting open-ended data from

interviewers describing their observational strategies and justifications for their judgments, rather than simply their *confidence* in the judgments. Analyses of these justifications (see Chapter 4) indicate that a relatively small proportion of justifications (12.1%) mention judgments based on guesses or "gut feelings," suggesting that interviewers are in fact able to articulate reasons for their judgments (contrary to Smith et al., 1991). Analyses of the associations of these reported justifications with judgment accuracy have the potential to provide insight into those strategies associated with reduced error (e.g., a focus on nonverbal cues, based on the "thin-slice" literature), and these analyses are presented in Chapter 4 of this dissertation.

Preliminary analyses of NSFG data have also indicated that a variety of paradata or "data collection process information" (Beaumont, 2005; Couper, 1998; Couper and Lyberg, 2005) collected by interviewers *prior* to screening interviews (e.g., number of attempted calls, previous statements made by potential respondents) and *during* the screening interviews (e.g., presence of a member of the opposite sex) are strongly predictive of eventual survey reports of sexual activity (West, 2010a). Design-based (i.e., taking the complex design features of the NSFG sample into account) logistic regression analyses of NSFG data collected from June 2006 to December 2008 (n = 13,495) indicated that rural areas, housing structures with many units, presence of children under 15, no previous establishment of contact, respondents making negative statements about the survey in immediate previous calls, selection of the informant as the respondent in the screening interview, older respondent age, presence of a member of the opposite sex, and multiple-adult households were all significant ($p < 0.05$) predictors of eventual respondent reports

of current sexual activity in the main NSFG interview. Interestingly, however, the bivariate correlation of the predicted probabilities of current sexual activity based on this model and the binary interviewer *judgments* of current sexual activity was only 0.33, suggesting that the interviewers may be picking up information about sexual activity that would not be predicted by a model. These analyses therefore suggest that information available to the NSFG interviewers at the time that they make a judgment about the sexual activity of a selected respondent (immediately after a completed screening interview) might be used to help the interviewers make a more informed (and potentially more accurate) guess.

This notion is supported by the psychological theory that "…a personality trait can be accurately judged if the judge can manage to detect and correctly use behaviors that are relevant to the trait and available to his or her observation" (Funder, 1995, p. 658). This theory is consistent with the theoretical framework described by Tversky and Kahneman (1974), which urged a focus on reliable and relevant predictors of a given phenomenon being judged. Similar theories can also be found in the psychological literature on intelligence analysis techniques employed by organizations like the Central Intelligence Agency (CIA). The *mosaic theory* of intelligence analysis posits that once all relevant and accurate pieces of information have been put together, a clear picture of reality emerges that leads to accurate estimates (Heuer, 1999, p. 62). Judgment accuracy, however, depends on both the accuracy of the judge's *mental model* (informed by which variables are most important for the judgment and how they relate to each other) and accuracy of the values attributed to variables included in the model (Heuer, 1999, p. 58-

59). This underscores the need for interviewers to be informed about appropriate models of characteristics being judged and to accurately measure the relevant variables in these models prior to making their judgments.

Preliminary testing of these theories was conducted in Quarter 15 of the NSFG (January 2010 to March 2010). Prior to the start of Quarter 15, all interviewers were informed by NSFG staff about the aforementioned predictors of sexual activity reports, and reminded of them each time that they made their judgments concerning sexual activity in the CAPI application. Specifically, interviewers received the following email at the beginning of Quarter 15 data collection:

Dear NSFG FRs (Field Researchers),

Below are the promised details about the Sexually-Active Observation modification. I've also included an answer to a question that came up on one of last week's conference calls and a note about STrak communications on December 27th.

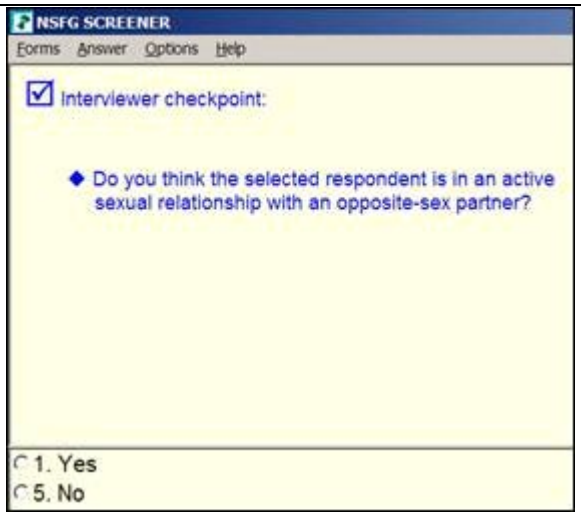
Note that there are 4 screenshots in this e-mail message. If you can't see them all please e-mail me ASAP. Thanks.

Have a wonderful break!

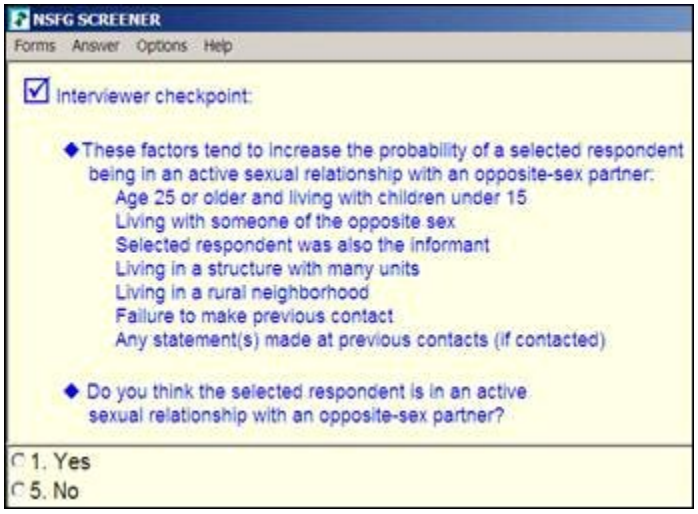
Shonda.

Screener Sexually-Active Observation Modifications

The question currently reads:



The new version will be:



Please contact Shonda via e-mail if you have any questions about the new version of the observation. Note that the factors now listed on the screen are included because they were identified via statistical modeling to increase the probability of the selected R being in an active sexual relationship with an opposite-sex partner. The model does not indicate *why* these factors might be important.

After these changes were implemented and data collection was completed in Quarter 15, cross-tabulations of interviewer judgments with actual survey reports of sexual activity

were examined, and compared with the same cross-tabulations in Quarter 14 (using a chi-square test for comparing Kappa statistics computed on independent samples). This analysis found evidence of significant changes in overall agreement between interviewer judgments and survey reports from Quarter 14 to Quarter 15 [Quarter 14 Kappa = 0.2466, 95% CI = (0.1813, 0.3120); Quarter 15 Kappa = 0.3660, 95% CI = (0.3035, 0.4284); $\chi^2_1 = 6.696$, $p < 0.01$][2] and a reduction in false positives (from 9.51% of observations to 6.67%). These preliminary results (West, 2010a) provide some initial support for the technique of supplying interviewers with relevant predictors of traits being judged to improve judgment accuracy. However, more in-depth analyses and randomized experiments are needed to control for interviewer effects and other factors that might be impacting accuracy when evaluating this type of intervention. The second study in this dissertation (Chapter 4) thoroughly evaluates the effectiveness of this NSFG intervention while controlling for potential confounders in a multilevel modeling framework.

The Implications of Errors in Interviewer Observations for Nonresponse Adjustments

Errors in interviewer observations are inevitable, and although certain techniques may prove to be effective at reducing error rates, they will never be eliminated completely. The implications of these errors for the effectiveness of nonresponse adjustments therefore require additional research, primarily to motivate the development of design

[2] There were 31 interviewers working in both Quarters 14 and 15 of the NSFG, introducing the possibility of correlated observations across the two quarters within these interviewers (despite the fact that information was collected on two independent random samples). Interviewer-specific Kappa statistics were computed for each of these 31 interviewers, and these measures of agreement did not have a significant correlation across the two quarters ($r = 0.1335$, $p = 0.4741$). The assumptions behind this chi-square test of equal Kappa coefficients for two independent samples therefore seem reasonable.

strategies for increasing the quality of the observations and statistical techniques for mitigating the effects of errors in the observations on nonresponse adjustments.

A small number of recent studies have suggested that errors in auxiliary variables will have a negative impact on the effectiveness of nonresponse adjustment techniques. Biemer et al. (2011) studied sources of error in interviewer-recorded numbers of call attempts in the National Survey of Drug Use and Health (NSDUH). Through simulation studies, these authors demonstrated that higher levels of *disposition-dependent* error (i.e., error rates that vary for refusals, completed interviews, etc.) in call records can substantially bias estimates of population proportions based on *callback models* (Biemer et al., 2010), when the different groups defining the proportion have different probabilities of responding. Despite finding that interviewer observations are strong predictors of both contact and cooperation in the ESS, Blom et al. (2011) speculate that errors and a lack of reliability in the interviewer observations may be resulting in under-estimation of these relationships. In another recent published example of this problem, Durrant et al. (2010, p. 22) speculate that errors in interviewer estimates of the ages of persons first contacted in UK household surveys was preventing unbiased estimation of interactions (hypothesized to exist based on liking theory; Groves, Cialdini, and Couper, 1992) between age of respondent and age of interviewer when predicting survey cooperation. Furthermore, Steiner et al. (2011) found that decreased reliability of covariates used in propensity score models for reducing the selection bias of treatment effects in observational studies will significantly impact the ability of covariates related to outcomes of interest to reduce selection bias. Collectively, these recent studies provide

motivation for examining the implications of higher error rates in interviewer observations for alternative nonresponse adjustments.

The statistical measurement error literature has firmly established the attenuating effects of errors in auxiliary variables on parameter estimates in both linear (Fuller, 1987) and logistic (Stefanski and Carroll, 1985) regression models. Analyses of NSFG data (see Chapter 3) have in fact demonstrated that the relationships of auxiliary variables with key survey variables are severely attenuated when estimating the relationships using interviewer judgments on the auxiliary variables (as opposed to the "true" measures of the auxiliary variables collected in the survey). The development of weighting classes for nonresponse adjustments (Little, 1986) that are homogeneous in terms of response propensities and survey variables of interest could therefore be adversely affected by the measurement error in these interviewer observations. Lessler and Kalsbeek (1992, Ch. 8, p. 190) describe mathematically how reduced within-class homogeneity in response propensities and survey variables can lead to an adjusted estimate of a mean that can have *more* bias than an *unadjusted* estimate of a mean, where the direction of the bias depends on the properties of the weighting classes. The reduced within-class homogeneity introduced by error-prone interviewer observations could therefore explain the lack of consistent effectiveness of nonresponse adjustments based on interviewer observations reported in the literature thus far. The first (Chapter 3) and third (Chapter 5) studies in this dissertation present simulations that examine this hypothesis in more detail.

The missing data mechanism underlying a given survey response process also needs to be considered when studying the impacts of error-prone auxiliary variables on nonresponse adjustments. If a missing data mechanism is such that nonresponse is indeed a function of values on the "true" auxiliary variable that interviewers are attempting to observe for respondents and nonrespondents, the survey collects information from respondents on this auxiliary variable, and the auxiliary variable is correlated with other key survey variables, then the missing data mechanism may in fact be non-ignorable (Little and Rubin, 2002). Standard weighting class adjustments or model-based imputation methods assuming ignorable missing at random (MAR) mechanisms may fail to reduce nonresponse bias as a result, especially when the interviewer judgments on the "true" auxiliary variable driving the nonresponse mechanism are prone to error and used in the adjustments. This could be another reason for the ineffectiveness of nonresponse adjustments based on interviewer observations reported thus far in the literature.

Pattern-mixture models (PMMs; Little, 1994; Little and Wang, 1996), which stratify incomplete data based on patterns of missing data and formulate distinct models for the variables within each stratum, have been shown to be effective likelihood-based tools for making unbiased inferences about parameters in the presence of a non-ignorable missing data mechanism. However, their utility in multipurpose surveys where auxiliary variables have been measured with error (as opposed to other more commonly used nonresponse adjustments) has not been evaluated in any studies to date. The third study in this dissertation (Chapter 5) develops PMM estimators for this setting, evaluates the

effectiveness of the PMM estimators in more detail, and contrasts their empirical performance with other commonly used estimators using simulations.

A Broader View of the Problem

In general, the quality of survey products depends heavily on careful approaches to survey research using a total survey error framework (see Groves and Lyberg, 2010). Auxiliary variables are used for a variety of purposes in producing the ultimate products of survey research, and the total survey error framework suggests that the error properties of auxiliary variables and the implications of these error properties for the quality of the products should not be ignored. A broader conceptual view of the problem of error in auxiliary variables suggests that other established survey estimation methodologies may also be affected by this problem.

First, there are several examples in the literature of samples being stratified based on auxiliary variables that are likely prone to error. Previous studies have stratified samples based on expert ratings of whether a household contains members of a rare population, according to publicly observable cultural indicators (Elliott et al., 2009); wealth rankings of housing units provided to interviewers by local informants (Adams et al., 1997); income rankings of housing units provided by interviewers (Hess, 2009, p. 8); estimated values of household socio-economic status (Apoyo Opinión y Mercado, 2004, p. 4); and interviewer judgments of general income levels for census blocks (e.g., EIA, 1997). Pickering and colleagues (2003) advocated that the EHCS use a *shadow sample* approach for selecting a sample that satisfied sample size requirements for key subgroups of

housing units. In this approach, features of the housing units in the shadow sample were either determined from interviews with neighboring houses in an ongoing survey, or from interviewer observations (if interviews from the ongoing survey were not completed), to maximize sample coverage. Excessive errors in the interviewer observations may have resulted in target sample sizes not being satisfied in important subgroups, but the authors argued that gains in coverage from using the observations outweighed reductions in the accuracy of these auxiliary variables from adding the observations.

Further, in the seventh cycle of the NSFG, second-phase samples of active NSFG nonrespondents after the first 10 weeks of a quarter, designed to reduce nonresponse bias (Groves and Heeringa, 2006), were stratified by response propensities predicted using interviewer observations (Lepkowski et al., 2010). Even more recently, researchers have studied the possibility of geographic stratification of cell phone samples based on the locations of switch centers and the purchase locations of cell phones, and found substantial errors in stratum codes at the county and sampling area levels (Wolter et al., 2011). In general, population units may also be misclassified into strata used for *post-stratification* adjustments. Kish (1965, p. 99) suggests that minor errors in stratification are not critical for the efficiency of estimates, but systematic evaluations of this problem in the literature are rare.

Many other survey methodologies aside from stratification use auxiliary variables that may be error-prone. Dual frame surveys for landline and cell phone users rely on respondent-reported counts of landlines and cell phones for estimation of selection

probabilities, and documented errors in these reports (e.g., Tucker et al., 2007) could bias sampling weights (Edwards et al., 2011; Merkle and Langer, 2008). In general, information about domain membership in dual frame surveys may need to be obtained directly from a respondent, and incorrect reports can lead to bias in dual frame estimators (e.g., Stokes and Lin, 2010). Multiplicity (or network) sampling of rare populations (Kalton and Anderson, 1986) relies on respondent-reported counts of related persons with rare traits, which could also bias selection probabilities. Interviewer observations on the perceived resistance of respondents toward the survey have been used to predict consent to physical measurement in a population of older adults (Sakshaug et al., 2010). The process of geocoding Census information to available addresses in Random Digit Dialing (RDD) samples can also lead to errors if the linked auxiliary information is mismatched to telephone exchanges (Biemer and Peytchev, 2010; Biemer and Peytchev, 2011).

Recent developments allowing survey researchers to link address-based sampling frames to purchased commercial databases (e.g., Experian) offer the promise of creating rich sampling frames (e.g., Fahimi, 2010), but the error properties of these linked auxiliary variables have received only minimal research focus (e.g., Daily et al., 2008; Hubbard and Lepkowski, 2009; Pickering et al., 2003, p. 8; Yan et al., 2011). Errors in auxiliary variables linked from administrative records have also been reported (e.g., Scioch and Bender, 2010). Recent work has developed methods for selecting the best vectors of fully observed auxiliary variables (including variables from administrative registers and process data) to be used in calibration estimators for reducing nonresponse bias (Sarndal and Lundstrom, 2010), but identification of these "best" vectors may be adversely

impacted by different levels of error in the auxiliary variables. Model-based approaches to imputation of item-missing data and subsequent multiple imputation approaches to inference (Little and Rubin, 2002) also rely on auxiliary variables for prediction (see Wagner, 2010, for an example using interviewer observations), and errors in these variables could bias imputations and the subsequent estimates. Finally, errors in auxiliary variables may also affect model-based approaches to finite population sampling and inference (Valliant et al., 2000), introducing bias and inefficiency in estimates (e.g., Bolfarine, 1991; West, 2010b).

As a supplement to the literature review presented in this chapter, the plethora of examples above suggests that the problem of errors in auxiliary variables has broad and important implications for the field of survey methodology. As initial steps in examining the magnitude of this problem, the studies presented in the following chapters focus on the errors in interviewer observations used for making nonresponse adjustments to survey estimates, and the implications of the errors for the quality of the adjustments.

Chapter III

The Quality and Utility of Interviewer Observations of Household Characteristics in the National Survey of Family Growth (NSFG)

Summary

Survey agencies on occasion have used interviewer observations collected on sample units to adjust survey estimates for nonresponse. Ideally, these observations are related to both response propensity and survey variables. Increasingly, survey organizations request that interviewers collect more observations to enhance nonresponse adjustments and improve survey efficiency. These observations are based on interviewer judgments about the characteristics of a unit, and as such are prone to error. Presenting analyses of data from the National Survey of Family Growth (NSFG) in the United States, this study examines the quality and statistical utility of one set of interviewer observations, and considers the implications of errors in the observations for the effectiveness of nonresponse adjustments.

Introduction

Given declining response rates in surveys of nearly all formats worldwide (Baruch and Holtom, 2008; Biener et al., 2004; Cull et al., 2005; Curtin et al., 2005; de Leeuw and de Heer, 2002; Tolonen et al., 2006) and rising costs of data collection, today's survey

researcher often relies on post-survey nonresponse adjustments to repair the nonresponse bias in survey estimates due to unit nonresponse. Many of these adjustments rely on auxiliary variables that are available for both the respondents and the nonrespondents in a given sample. Reductions of both the bias and variance in estimates that can arise from unit nonresponse are possible when these auxiliary variables are related to both the survey variables of interest and response indicators (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Kreuter et al., 2010; Lessler and Kalsbeek, 1992; Little and Vartivarian, 2005). Unfortunately, the number of auxiliary variables having these properties is small in practice (Kreuter et al., 2010).

Due to improvements in data retrieval techniques, large survey research programs have started to examine paradata (Beaumont, 2005; Couper, 1998; Couper and Lyberg, 2005), or variables that measure the survey data collection process, including interviewer observations (Kreuter et al., 2010). A growing body of research has found associations of these paradata, and particularly interviewer observations, with establishment of contact with sampled households (Blom et al., 2011; Durrant et al., 2011b), response indicators (Blom, 2009; Blom et al., 2011; Campanelli et al., 1997, Section 4.6; Casas-Cordero, 2010a; Durrant et al., 2011a; Durrant et al., 2010, p. 13; Durrant and Steele, 2009; Groves and Heeringa, 2006; Lynn, 2003; Maitland et al., 2009; Scholes et al., 2008), and key survey variables (Casas-Cordero, 2010a; Gonzalez and Kreuter, 2010; Kreuter et al., 2010; Maitland et al., 2009; Scholes et al., 2008). Interviewer observations are typically judgments, making them prone to error. Few published studies have examined these errors in judgment, and the implications of these judgment errors for the effectiveness of

post-survey nonresponse adjustments have not yet been examined. This study aims to address this gap in the literature.

Data are taken from the seventh cycle of the National Survey of Family Growth (NSFG). The NSFG has used paradata extensively for production and estimation work (Groves et al., 2009, p. 23-24). Beginning in Cycle 7 of the NSFG (June 2006 to June 2010), interviewers were requested to estimate whether children under the age of 15 were present in a sampled household prior to an eligibility screening interview. In addition, immediately after the completion of the screening interview, interviewers were requested (on the doorstep) to judge whether or not a selected respondent (between the ages of 15 and 44) was in a sexually active relationship with a member of the opposite sex. These two observations were made for both responding and nonresponding households and sample persons. The observations were collected as part of a larger responsive survey design (Groves and Heeringa, 2006), and intended for use in predicting response propensities for individuals and nonresponse adjustment of survey weights. The two variables have the important property of being theoretically correlated with several key NSFG variables.

This investigation seeks to address four questions regarding these interviewer observations in the NSFG:

1. What are the error properties of the interviewer observations?
2. Are the observations associated with a) response indicators and b) key survey variables?

3. Do key survey estimates shift when using the observations for nonresponse adjustments?
4. How do errors in the observations affect nonresponse adjustments?

Background

Previous work has examined the use of paradata for nonresponse adjustments. For example, Peytchev and Olson (2007) used similar types of paradata for making nonresponse adjustments in the American National Election Studies, while Kreuter, Lemay, and Casas-Cordero (2007) included paradata in nonresponse adjustments for the European Social Survey. Yan and Raghunathan (2007) in a national transportation survey, Gonzalez and Kreuter (2010) in the Consumer Expenditure Survey, Blom (2009) in the European Social Survey, and Scholes et al. (2008) in the British Gambling Prevalence Survey also used paradata in developing nonresponse adjustment weights.

A recent study by Kreuter et al. (2010) demonstrated that the NSFG interviewer observations are stronger correlates of key NSFG variables than similar paradata collected in other surveys. Kreuter et al. (2010) also showed that incorporating the observations into nonresponse adjustments led to moderate shifts in NSFG estimates.

The associations of interviewer observations with response indicators and key survey variables may be attenuated by errors in the observations. Such attenuation could in turn reduce the effectiveness of nonresponse adjustments based in part on the observations. The impact of error in auxiliary variables on the bias of estimated regression coefficients

in linear regression models (e.g., Fuller, 1987) and logistic regression models (Stefanski and Carroll, 1985), which are often used for making nonresponse adjustments based on predicted response propensities, has been well-established. The homogeneity of weighting classes constructed for nonresponse adjustments (in terms of response indicators and key survey variables) could therefore be adversely affected by the errors in these observations.

For example, Lessler and Kalsbeek (1992, Ch. 8, p. 189-190, Equations 8.18 and 8.20) show that when the true respondent mean on a variable of interest tends to be higher than the true nonrespondent mean within each class, and the true respondent means and the expected response rates are inversely related among all classes, an adjusted mean will have *more* bias than an unadjusted mean. "…These results warn us that it is possible to do more harm than good by using weighting class adjustments," (Lessler and Kalsbeek, 1992, p. 190). Errors in the interviewer observations used to form the classes could lead to just such an outcome.

To date, a few studies have *directly* examined the errors in interviewer observations, such as Groves et al. (2007), Pickering et al. (2003), Sturgis and Campanelli (1998), and Tipping and Sinibaldi (2010). These studies examined interviewer observations that could be validated, and found that accuracy rates ranged from 46% to 92% (Groves et al., 2007; McCulloch et al., 2010; Pickering et al., 2003; Sturgis and Campanelli, 1998; Tipping and Sinibaldi, 2010). Pickering et al. (2003) suggested that simulations be used to examine the implications of inaccuracies in observations for sampling purposes. Hahn

et al. (1996) and Drury et al. (1980) also report frequent inaccuracies in interviewer observations of self-reported ethnic classifications, particularly for ethnic minorities.

Other existing work has examined *indirect* indicators of error in the interviewer observations. Sinibaldi (2010) examined consistency in both objective and subjective observations over time. Casas-Cordero (2010b) investigated within-area correlation in observations as another measure of accuracy, and Feldman (1951) examined interviewer variance in subjective judgments using interpenetrated assignments. Pickering et al. (2003) reviewed interviewer problems with making the observations, and Kreuter et al. (2007) and Lynn (2003) examined missing data rates for the observations as a method to assess the quality of the observations. Alwin (2008, p. 151) and Mosteller (1944) indicate that the inter-interviewer reliability of household observations tends to be higher than that of judgments about *respondent* features. Kennickell (2011, p. 6) indicates that the Survey of Consumer Finances (SCF) actually ceased collection of interviewer observations due to lower inter-interviewer reliability.

No studies have specifically considered the implications of errors in interviewer observations for the effectiveness of nonresponse adjustments based in part on the observations. Steiner et al. (2011) considered the context of observational studies, where propensity score modeling is often used to reduce possible selection bias in treatment effects. These authors demonstrated that when the covariates used in propensity score models are related to outcomes of interest but have reduced reliability, the ability of the covariates to reduce selection bias will be significantly reduced. Biemer et al. (2011)

studied sources of error in interviewer-recorded numbers of call attempts in the National Survey of Drug Use and Health (NSDUH), and the bias that these errors can introduce in nonresponse adjustments based on a *callback model* (Biemer et al., 2010). These authors showed through simulation studies that when error rates for the call records vary depending on the disposition of a sample case (e.g., refusals have higher error rates compared to completed cases), the bias in estimated proportions based on the callback models can be substantial, particularly when there is differential response propensity across groups with high and low levels of the estimated proportion.

Given the increasing interest in using these observations in nonresponse adjustment, the present study aims to address an important gap in the literature about the potential losses in the effectiveness of nonresponse adjustments due to interviewer observation error.

NSFG Data

The target population of the NSFG is non-institutionalized males and females between the ages of 15 and 44 living in the United States. The survey has a primary goal of collecting "nationally representative data on factors affecting birth and pregnancy rates, family formation, and the risks of HIV and other STDs" (Groves et al., 2009).

NSFG interviewing is conducted in two steps. First, a screening interview is conducted with cooperating sample households to determine whether anyone in the household is eligible for the study (the screener interview). This initial screening interview is then followed by a main interview about births, pregnancies, and other family growth topics

with a randomly selected eligible person in the household. All screening and main interviews are conducted face-to-face by trained female interviewers using a Computer Assisted Personal Interviewing (CAPI) application on a laptop computer. For more sensitive questions in the main interview, Audio Computer-Assisted Self Interviewing (A-CASI) is used, and the interviewer gives the respondent the laptop and a pair of headphones so that all sensitive questions can be self-administered.

The seventh cycle of the NSFG also had two phases of data collection. In the first 10 weeks of a data collection quarter (Phase 1), standard data collection procedures were applied to all sample cases. In the last two weeks of a quarter (Phase 2), a more intensive data collection protocol was applied to a subsample of active nonrespondents (to both screener and main interview requests), in an effort to maximize response rates for the quarter. Cases with a higher propensity of responding (based on a response propensity model that included various interviewer observations as covariates) were subsampled at a higher rate in the second phase.

NSFG interviewers are typically assigned to work in a single primary sampling unit (PSU). Prior to the first face-to-face contact attempt with a randomly selected household, interviewers first locate the household and then record a judgment of whether the selected household contains any children under the age of 15 (yes / no). There were a total of 58,225 pre-screener observations on the presence of young children reported by 116 interviewers from Cycle 7. For each, completed household roster information was available to determine whether children under the age of 15 were actually present in the

household. (Observations with missing household roster information were deleted for the purposes of this study.) We assume completed household rosters are correct. Interviewers also noted physical impediments to the household, safety concerns in the area where the household was located, and whether all households in the area were residential.

Immediately after the completion of the screening interview and the selection of an eligible person for the main interview, interviewers were asked to judge whether the selected person was in a sexually active relationship with an opposite-sex partner (yes / no). There were a total of 22,669 judgments of sexual activity by 113 interviewers where actual survey information on sexual activity was also available from the *completed* main interview. It was again assumed that the survey information was correct, for comparison with the interviewer observation. Since the "true" values for current sexual activity were extracted from the main interview, analyses of the relationship of the "true" current sexual activity with a response indicator are not possible for the present investigation. After the completion of a screening interview, interviewers were also asked to estimate the probability that a main interview would be completed (high, medium, or low).

There were 25,451 completed screening interviews where the two interviewer observations of interest (presence of young children and current sexual activity) were available for potential respondents *and* nonrespondents. Cases with completed screening interviews that were not randomly sampled for a second phase follow-up of active nonrespondents were excluded from this data set. Main NSFG interviews were completed by a total of 22,682 respondents. Thirteen (13) of these respondents had missing data on

the variables necessary to determine a reported value of current sexual activity, resulting in 22,669 sample persons with sufficient data for studying the error in these interviewer judgments. Sampling weights and sampling error codes provided by NSFG staff were used for unbiased estimation of selected parameters and design-based estimation of the standard errors of the parameter estimates (Lepkowski et al., 2010).

Several survey variables collected in the main NSFG interview were examined in this study. These included 1) a binary indicator of whether the respondent had never been married; 2) a binary indicator of whether the respondent had ever cohabitated with a partner; 3) the number of sexual partners in the past year [reported in the Audio Computer Assisted Self Interview (ACASI) version of the main interview]; 4) for males, a count of biological children; and 5) for females, parity, or the number of live births. Male and female respondents to the main interview were coded as being sexually active if reporting one or more opposite-sex partners in the past 12 months. Female respondents were also asked about having a current opposite-sex partner, and this measure was used to indicate being sexually active if no information was available on the number of partners in the past 12 months. Measurement error was certainly possible for these variables, and error rates may have differed for males and females, but implications of these errors are left to future research.

Statistical Analyses

Four separate statistical analyses were performed to address the first three research questions. For research question 1, unweighted Kappa statistics were used to examine overall agreement of the interviewer judgments with survey measures.

For research question 2, mixed-effects logistic regression models were used to regress an indicator of response to the main interview (conditional on a completed screening interview) on a series of auxiliary variables identified as important in previous response propensity models for the main NSFG interview (Lepkowski et al., 2006). In addition, the base NSFG sampling weight, which adjusts for unequal probabilities of selection, was included as a covariate in order to make the sampling uninformative with respect to the model. Random interviewer effects were included in the models to adjust for the potential clustering of response indicators within interviewers (that is, interviewers having higher and lower levels of response propensity). These models were fit for the 25,451 successful screening interviews with interviewer observations available.

Three versions of this response propensity model were estimated. The first did not include any predictors representing interviewer observations. The second added interviewer observations that could *not* be validated as predictors (noting physical impediments to the household, estimated probabilities of a main interview being completed, noting whether all housing units in a segment are residential, and noting safety concerns). The purpose of the second model was to assess the additional contribution made to the propensity model by these interviewer observations. The third model then added the two interviewer judgments of current sexual activity and presence

of young children as predictors, to analyze their independent ability to predict response to the main interview. These three models were also estimated treating the interviewer effects as fixed, enabling comparisons of re-scaled pseudo R-squared values for the three models (Nagelkerke, 1991).

The three models were fitted using residual pseudo likelihood estimation (SAS PROC GLIMMIX), and the null hypothesis that the variance of the random interviewer effects was zero was tested using a likelihood ratio test (Zhang and Lin, 2008). Predicted response propensities were computed for each responding case, as were the empirical best linear unbiased predictions (EBLUPs) of the random interviewer effects. To minimize additional variance in the estimates introduced by variance in the nonresponse adjustments, these response propensities were grouped into deciles (or weighting classes), and the inverse of the unweighted proportion of responding cases within each of the 10 weighting classes was used to adjust the base sampling weights for nonresponse (Little, 1986).

Also for research question 2, mixed-effects logistic and linear regression models were estimated by regressing the five survey variables for NSFG respondents only on a number of auxiliary variables. Count responses were transformed using square-root transformations to stabilize variance in the responses (Faraway, 2005, p. 58). Predictor variables included the two interviewer observations that could be validated, along with the same base sampling weight and other interviewer observations used above. Random interviewer effects were also included as in the response propensity model. These models

could only be fitted using survey respondent data, and thus no inferences can be made to nonrespondents as well unless it is assumed that the associations of the survey variables with the interviewer observations are the same for both respondents and nonrespondents (Maitland et al., 2009; Peytcheva and Groves, 2009). Differential error rates for the observations for respondents and nonrespondents (e.g., Matsuo et al., 2010, p. 35-43) may lead to different associations, and this has implications for nonresponse adjustments (e.g., Biemer et al., 2011); unfortunately, the available NSFG data did not permit testing this assumption, because only the respondents had information on both the interviewer observations and the survey variables available.

For research question 3, design-based estimates of means and percentages on the five survey variables were computed using three different weights: 1) the base sampling weights that adjusted for unequal probabilities of selection; 2) nonresponse-adjusted base weights with weighting class adjustments *excluding* the two interviewer judgments under study; and 3) nonresponse-adjusted base weights with weighting class adjustments *including* the two interviewer judgments. The nonresponse adjustments assume that the NSFG nonrespondents are missing at random (Little and Rubin, 2002). Variances and covariances of these estimates were estimated using Taylor Series Linearization (Wolter, 2007) and account for the complex sample design of the NSFG, but ignore negligible finite population corrections (based on the large NSFG target population). Weighted estimates were compared by "stacking" three different versions of the same data file (with only the weights differing), and constructing confidence intervals for the differences in the means that incorporated the covariances of the estimates. These

analyses were performed using the `svy: mean, over()` command in conjunction with the `lincom` post-estimation command in Stata 11.2.

Finally, a small simulation study was performed using real NSFG data to address the fourth research question. A hypothetical population included $N = 10{,}561$ female respondents to the main NSFG interview (Cycle 7) with complete data on selected variables. These variables included the interviewer judgment about current sexual activity for the selected subject and the actual reports of current sexual activity from the main NSFG interview. Parity and number of partners in the past year were also included.

In each of six simulations (three weighting schemes for each survey variable), 1,000 simple random samples of size $n = 500$ were selected from the population. Unit nonresponse was simulated for each of the 1,000 samples using the following logistic regression model (based on actual NSFG outcomes; see Table 3.3):

$$\Pr(response_i) = \frac{\exp(report.sexually.active_i)}{1 + \exp(report.sexually.active_i)}$$

A sampled case denoted by i had values on the two survey variables deleted if a random draw from a *UNIFORM*(0,1) distribution was greater than or equal to the probability computed above. The simulated probability of response was thus a function of the *reported* sexual activity for case i (1 = yes, 0 = no), and not the interviewer judgment.

In four of the six simulations, a simple weighting class adjustment for nonresponse was performed. Two weighting classes defined by categories for either the "true" respondent report of current sexual activity or the interviewer judgment were formed, and the inverse

of the proportion of respondents within each class was used as a nonresponse adjustment weight. Two simulations computed nonresponse-adjusted estimates of mean parity for each sample using the two alternative auxiliary variables, while two simulations computed nonresponse-adjusted estimates of mean number of partners in the past year using the alternative adjustments. Finally, two simulations computed complete case estimates of means for parity and number of partners in the past year for each sample.

Since the means for the two survey variables are known, the following outcomes were examined for each of the six simulations: 1) the empirical bias of the estimate (in terms of a percentage bias relative to the known mean); 2) the empirical root mean squared error of the estimate (RMSE), or the square root of the sum of the empirical bias squared and the variance (MSE); and 3) 95% confidence interval coverage of the estimate. Standard errors of weighted estimates were computed using Taylor Series Linearization, and confidence intervals were computed assuming normally distributed estimates.

Results

What are the error properties of the interviewer observations in the NSFG?

Table 3.1 shows that roughly 72.3% (i.e., 59.94% + 12.36%) of the interviewer judgments on presence of young children were accurate, i.e., in agreement with the survey data. The Kappa statistic was 0.285 (95% CI = 0.276, 0.293), a level of agreement considered 'fair' per Landis and Koch (1977). The false positive rate was 0.169 (7,103 / 42,001), while the false negative rate was 0.557 (9,028 / 16,224), indicating that

interviewers were much more likely to judge that households did not have young children when in fact they did.

Table 3.1: Case counts and overall percentages indicating the error properties of interviewer judgments regarding the presence of children under the age of 15 in selected households (NSFG, Cycle 7).

	Household Roster Indicator: Kids Age < 15		
Interviewer Judgment: Kids Age < 15	No	Yes	Totals
No	**34,898 (59.94%)**	9,028 (15.51%)	43,926 (75.44%)
Yes	7,103 (12.20%)	**7,196 (12.36%)**	14,299 (24.56%)
Totals	42,001 (72.14%)	16,224 (27.86%)	58,225 (100.00%)

NOTE: Kappa Statistic = 0.285, 95% CI = (0.276, 0.293).

Interviewers had an easier time judging sexual activity (Table 3.2), with overall accuracy (for main interview respondents only) approaching 78%. The Kappa of 0.334 (95% CI = 0.319, 0.349) for Table 3.2 also indicates 'fair' agreement, as in Table 3.1.

Table 3.2: Case counts and overall percentages indicating the error properties of interviewer judgments of current sexual activity among respondents (NSFG, Cycle 7).

	Main NSFG Interview: Selected R Sexually Active		
Interviewer Judgment: Selected R Sexually Active	No	Yes	Totals
No	**2,230 (9.84%)**	2,081 (9.18%)	4,311 (19.02%)
Yes	2,912 (12.85%)	**15,446 (68.14%)**	18,358 (80.98%)
Totals	5,142 (22.68%)	17,527 (77.32%)	22,669 (100.00%)

NOTE: Kappa Statistic = 0.334, 95% CI = (0.319, 0.349).

In contrast to the housing unit observations on presence of young children, the false positive rate for the sexual activity judgments was much higher (0.566) than the false negative rate (0.119). This indicates that interviewers' judgment of sexual activity was

much more difficult for sample persons who were not sexually active (or that interviewers may simply default to the modal value for current sexual activity).

Are the observations associated with response indicators and key NSFG variables?

Table 3.3 presents estimates of adjusted odds ratios (along with 95% confidence intervals for the odds ratios) in the three logistic regression models predicting propensity to respond to the main NSFG interview. Estimated odds ratios for additional auxiliary variables (base sampling weight, call number, number of contacts, black respondent, indicators for quarters 1-15, age of selected respondent, urban primary sampling unit, single-person household, non-white interviewer, bilingual interviewer, Census regions, sampling segment domains, and second phase sample indicator) are not shown to simplify the presentation, but are available upon request.

Table 3.3: Selected main interview response propensity modeling results, showing adjusted relationships of NSFG interviewer observations with response indicators.

	Model 1	Model 2	Model 3
Interviewer Observation	Estimated OR (95% CI)	Estimated OR (95% CI)	Estimated OR (95% CI)
Interviewer Notes Physical Impediments to Household		0.985 (0.851, 1.139)	0.986 (0.852, 1.142)
Interviewer Estimates High Main Interview Probability		0.565 (0.479, 0.668)	0.559 (0.473, 0.661)
Interviewer Estimates Medium Main Interview Probability		0.326 (0.260, 0.341)	0.295 (0.257, 0.339)
Interviewer Estimates Low Main Interview Probability		0.093 (0.081, 0.106)	0.093 (0.081, 0.106)
Interviewer Does Not Report Main Interview Probability		Reference	Reference
Interviewer Notes All Housing Units in Segment Residential		0.997 (0.902, 1.101)	0.999 (0.904, 1.104)
Interviewer Notes Safety Concerns		1.026 (0.915, 1.151)	1.018 (0.907, 1.142)
Interviewer Estimates Respondent Sexually Active			1.923 (1.707, 2.166)
Interviewer Estimates Children Under 15 in Household			1.184 (1.064, 1.317)
Sample Size	25,451	25,451	25,451
Estimated Variance of Random Interviewer	0.261 ($p < 0.001$)	0.257 ($p < 0.001$)	0.274 ($p < 0.001$)

Effects (LRT p-value)			
Generalized Chi-square Statistic (SAS, 2011)	24,763.98	23,005.99	22,327.04
Pseudo R^2 (Treating Interviewer Effects as Fixed)	0.271	0.357	0.365

NOTES: OR = Odds Ratio, LRT = Likelihood Ratio Test, CI = Confidence Interval. Estimated odds ratios for additional auxiliary variables (base sampling weight, call number, number of contacts, black respondent, indicators for quarters 1-15, age of selected respondent, urban primary sampling unit, single-person household, non-white interviewer, bilingual interviewer, Census regions, sampling segment domains, and second phase sample indicator) are not shown and are available upon request. Models including fixed interviewer effects (enabling computation of pseudo R-squared values) did not include the interviewer-level covariates (non-white interviewer and bilingual interviewer) to avoid confounding.

The results in Table 3.3 indicate that the interviewer observations that could not be validated (noting physical impediments to the household, estimated probabilities of a main interview being completed, noting whether all housing units in a segment are residential, and noting safety concerns) make a substantial contribution to the NSFG response propensity model. There is both a sizable reduction in the generalized chi-square fit statistic and an increase in the pseudo R-squared value when interviewer observations are added to Model 1. Specifically, interviewer estimates of the probability that a main interview will be completed for a completed screener are strongly predictive of response indicators, with cases having missing estimates (many of which are missing due to main interviews that are completed immediately following the screening interview) and cases having high predicted probabilities having relatively higher propensities to respond.

The two interviewer judgments that could be validated are also significant predictors of response (Model 3), but they do not result in the same large improvement in model fit. Households estimated to have children under 15 and selected respondents estimated to be sexually active had significantly higher probabilities of completing the main NSFG

interview. We cannot examine the contributions of the "true" auxiliary variables measuring the presence of young children and current sexual activity to the response propensity model, since these variables were only both available for main interview respondents. There is evidence of substantial interviewer variance in response propensities, even after accounting for interviewer-level predictors of being non-white or bilingual, indicating that predicted response propensities should account for EBLUPs of the random interviewer effects in these models.

Table 3.4 displays, for respondents only, adjusted estimates of the relationships of the two interviewer observations with the five key NSFG variables. The estimated coefficients are taken from models that included random interviewer effects and all of the aforementioned auxiliary variables as predictors (base sampling weight, call number, number of contacts, black respondent, indicators for quarters 1-15, age of selected respondent, urban primary sampling unit, single-person household, non-white interviewer, bilingual interviewer, Census regions, sampling segment domains, and second phase sample indicator). Two estimates are presented for each judgment: the estimated coefficient for the judgment, and the estimated coefficient for the 'correct' survey variable.

Table 3.4: Adjusted estimates of regression parameters for the two interviewer judgments as predictors of five key NSFG variables, contrasted with estimates using the "true" auxiliary variables as predictors instead (NSFG respondents only).

NSFG Variable	n	Pseudo R^2 Values	Interviewer Judgment: Children Under 15	Household Roster: Children Under 15	Interviewer Judgment: Sexual Activity	Respondent Report: Sexual Activity
Never Been Married	22,682	0.523 / 0.571	-0.20***	-0.80***	-1.17***	-1.95***
Ever Cohabitated	22,682	0.283 / 0.364	0.13***	0.32***	0.98***	1.99***

Number of Biological Children (males only)	10,403	0.369 / 0.466	0.20***	0.48***	0.22***	0.43***
Number of Sexual Partners in Past Year	21,008	0.054 / 0.605	-0.01	-0.05***	0.26***	1.17***
Parity (# live births) (females only)	12,279	0.418 / 0.489	0.22***	0.73***	0.23***	0.32***

NOTES: Parameter estimates for other covariates listed in the Table 3.3 notes are not shown for each dependent variable in the first column. Pseudo R-squared values are computed from models with interviewer effects treated as fixed and interviewer-level covariates omitted (as in Table 3.3). The first pseudo R-squared value is for the model using the interviewer judgments, while the second value is for the model using the true values. Parameter estimates and tests of significance are based on models including random interviewer effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The interviewer observations tended to have strong associations with the five key survey variables when adjusting for the other auxiliary variables, including the base sampling weights (which were significantly related to three of the five survey variables, indicating informative sampling). However, it would appear that errors in the observations are severely attenuating the relationships of the two "true" auxiliary variables with the survey variables. Compare columns 4 and 5 and columns 6 and 7 in Table 3.4: the coefficients for the interviewer judgments are substantially smaller than the coefficients for the true survey values. There are also large differences in the R-squared or pseudo R-squared values for the fitted models when using the interviewer judgments rather than the "true" survey values, especially in the case of number of sexual partners in the past year. This suggests that use of the interviewer judgments may limit the collective predictive power of these auxiliary interviewer judgment variables, thus limiting the effectiveness of nonresponse adjustments including these two auxiliary variables.

Finally, likelihood ratio tests for the variances of the random interviewer effects in these ten models (not shown in Table 3.4) were significant for nine of the ten models (with the exception of the model including true values of current sexual activity). This suggests that there is substantial interviewer variance in responses even after accounting for a large number of auxiliary variables, including interviewer-level covariates.

Do key estimates shift when using the observations to adjust for nonresponse?

Table 3.5 presents estimates and design-based estimates of standard errors on the five key NSFG variables using the three alternative weights.

Table 3.5: Impacts of alternative nonresponse adjustments on NSFG estimates (linearized standard errors reported in parentheses).

Estimate (Sample Size)	**Base Weights Only**	**Nonresponse Adjusted Base Weights Without the Two Judgments**	**Nonresponse Adjusted Base Weights With the Two Judgments**
% Never Married (n = 22,682)	50.473% (1.002)	49.995% (1.179)	50.078% (1.203)
% Ever Cohabitated (n = 22,682)	49.533% (1.314)	49.881% (1.449)	49.782% (1.473)
Males: Mean # Biological Kids (n = 10,403)	1.266 (0.054)	1.304 (0.083)	1.300 (0.083)
Mean # Partners in Past Year (n = 21,008)	1.130 (0.015)	1.130 (0.016)	1.128 (0.016)
Females: Mean Parity (n = 12,279)	1.293 (0.041)	1.280 (0.043)	1.276 (0.044)

NOTE: These estimates do not incorporate post-stratification factors and do not represent final estimates based on NSFG Cycle 7.

The largest shifts from estimates computed using the base weights only to estimates computed with nonresponse-adjusted base weights occur when using other auxiliary variables and interviewer observations (that could not be validated) to form the weighting classes. Adding the two interviewer judgments that could be validated to the response

propensity models caused slight shifts in the estimates, which is consistent with the slight changes in the fit of the response propensity model when adding these judgments as predictors.

Notably, the nonresponse-adjusted mean for parity was found to be significantly lower than both the mean based on base weights only (95% CI for Difference = 0.0002, 0.0335) and the nonresponse-adjusted mean excluding the two judgments (95% CI for Difference = 0.0001, 0.0078). This suggests that including these two interviewer judgments in the nonresponse adjustment did have a significant (although small) impact on this particular estimate. In addition, the difference between the two adjusted means for number of partners in the past year was found to approach significance (95% CI = -0.0004, 0.0037), suggesting that the two observations were also causing a slight shift in this estimate. None of the other shifts were found to be significant.

How do errors in the observations impact nonresponse adjustments?

Table 3.6 summarizes the results of the small simulation study designed to compare the empirical performance of nonresponse adjustments based on actual survey reports of current sexual activity with that of adjustments based on interviewer judgments about current sexual activity.

Table 3.6: Results of simulation study, showing empirical performance of estimators with nonresponse adjustments based on either respondent reports of current sexual activity or interviewer judgments of current sexual activity.

NSFG Variable	Nonresponse Adjustment Method	Auxiliary Variable for Nonresponse Adjustment	True Mean	Empirical Bias (Rel. %)	Empirical RMSE	95% CI Coverage	Mean CI Width
Parity	None		1.3743	0.0064	0.0755	0.953	0.2975

				(0.47%)			
	Weighting Classes	Self-reported Sexual Activity	1.3743	-0.0027 (-0.20%)	0.0762	0.948	0.2986
		Interviewer Judgment of Sexual Activity	1.3743	0.0085 (0.62%)	0.0714	0.968	0.2975
Partners in Past Year	None		1.1626	0.0319 (2.67%)	0.0566	0.904	0.1788
	Weighting Classes	Self-reported Sexual Activity	1.1626	0.0012 (0.10%)	0.0453	0.960	0.1850
		Interviewer Judgment of Sexual Activity	1.1626	0.0326 (2.80%)	0.0573	0.906	0.1803

NOTE: CI = confidence interval.

The results in Table 3.6 show that the use of interviewer judgments of current sexual activity as an auxiliary variable when constructing the nonresponse adjustments (rows 3 and 6) attenuates potential reductions in bias when using the weighting class adjustment method relative to adjustments using the "true" self-reported values of current sexual activity (rows 2 and 5). This is especially true for survey variables having a stronger relationship with the auxiliary variable in question. In the hypothetical population, the correlation of the "true" current sexual activity with parity was only 0.072, while the correlation of "true" current sexual activity with the number of partners in the past year was 0.396. Table 3.6 shows that the adjusted estimate of the mean number of partners in the past year is the one that is most severely affected by using the interviewer judgments to define the weighting classes.

When using the interviewer judgments of current sexual activity to define the two weighting classes, the bias of the resulting estimates is similar to that found when

analyzing the complete cases (rows 1 and 4). The positive bias in the complete case estimator for the mean number of partners in the past year (2.67% relative bias) actually becomes *larger* when using nonresponse adjustments based on the interviewer judgment (2.80% relative bias). This result is consistent with the theoretical possibility suggested by Lessler and Kalsbeek (1992), and occurs here because a) respondents tended to have a higher mean number of partners in the past year than non-respondents in the two classes formed by the interviewer judgment, b) the mean number of partners in the past year was actually *higher* in the weighting class of those judged to not be sexually active by the interviewers, and c) response propensity was lower in the weighting class of those judged not to be sexually active by the interviewers.

Considering the estimators of mean number of partners in the past year, there is evidence of a higher empirical RMSE in the nonresponse adjusted estimator (compared to the complete case estimator) that uses the interviewer judgments to form the weighting classes. The RMSEs in Table 3.6 indicate that the variances of the three estimates for parity are 0.0057, 0.0058, and 0.0050, respectively, while the variances of the three estimates for number of sexual partners are 0.0022, 0.0021, and 0.0022, respectively. These results suggest that differences in the RMSEs of the estimators are largely being driven by the bias introduced by the interviewer judgments, rather than the variance.

There is also evidence in Table 3.6 that confidence interval coverage may be affected by the use of the interviewer judgments in nonresponse adjustments. This is especially true for variables having a stronger relationship with the interviewer judgments.

Discussion

This study addressed four research questions regarding the quality and utility of interviewer observations in the NSFG.

For the first research question, the quality of two interviewer judgments that could be validated was largely consistent with findings in the extant literature. Accuracy rates fell between 70 and 80 percent. The present study builds on previous work by showing that errors in specific observations may be systematic rather than variable.

For the second research question, the two judgments were found to have significant associations with both a main interview response indicator and five key NSFG variables, even when adjusting for a variety of other auxiliary variables and interviewer observations. These relationships remained significant when not adjusting for the other auxiliary variables as well. However, relationships of the two "true" auxiliary variables with the five survey variables were severely attenuated by errors in the judgments. These results suggest that interviewer judgments may be useful for nonresponse adjustment purposes, but need to be improved in accuracy.

For the third research question, weighting class adjustments incorporating the two interviewer judgments shifted estimates slightly relative to adjustments without the judgments, with significant or marginally significant shifts observed for estimates of the mean parity (for females only) and the mean number of partners in the past year. These

findings were consistent with existing literature showing that nonresponse adjustments incorporating interviewer judgments do not tend to shift estimates substantially.

Finally, a fourth research investigation mounted a simulation to assess the effect of the errors in one of the NSFG judgments on nonresponse adjustment weights. This was merited because associations of the "true" auxiliary variables with key NSFG variables were shown to be severely attenuated when using the interviewer judgments in their place. The simulation study confirmed that the use of interviewer observations instead of true values on auxiliary variables to form weighting classes for nonresponse adjustments can attenuate potential reductions in bias, and, in the case of auxiliary variables having stronger relationships with key survey variables, lead to more bias in estimates than complete case analyses. Importantly, this simulation study only considered total sample estimates in the NSFG; a great deal of survey research focuses on subpopulation estimates. The weighting classes considered in the simulations will likely become more homogeneous when focusing on specific NSFG subpopulations, especially if a given subpopulation variable is associated with the variables studied here. The bias of adjusted estimates based on the interviewer judgments would therefore be expected to fall in-between that of complete case estimates (high) and that of adjusted estimates based on true values (low).

These findings have many implications for survey practice. Interviewer observations clearly have the potential to be effective auxiliary variables for use in nonresponse adjustments, but error levels may limit this potential. Survey managers and survey

statisticians need to consider strategies for improving the quality of the judgments, especially given the systematic nature of the errors found in this study, and estimation techniques that mitigate the effects of the errors in the interviewer observations on nonresponse adjustments. For example, West (2010) showed that providing interviewers with observable predictors of eventual sexual activity reports helped to reduce false positive rates in their judgments, and Chapter 4 expands on this finding. Future studies might consider replacing particular judgments with "smoothed" predictions based on predictors of eventual respondent reports that are available at the time of an observation.

Survey statisticians could use a variety of estimation techniques to mitigate the effects of errors in interviewer observations on nonresponse adjustments. These include Monte Carlo EM methodologies for model estimation based on partitioned likelihoods, including models for the errors in covariates (Yi et al., 2011), misclassification simulation-extrapolation (MC-SIMEX) for binary auxiliary variables measured with error in logistic regression models (Küchenhoff et al., 2006), nonparametric maximum likelihood estimation methods that relax assumptions of normality for true covariates and the measurement errors (Rabe-Hesketh et al., 2003), pattern-mixture modeling (Little, 1994; Little and Wang, 1996; West and Little, 2012; see also Chapter 5 below), and regression calibration (Rosner et al., 1990; Spiegelman et al., 2000). Importantly, these methods all require access to validation data for the error-prone auxiliary variables. Kott (2006) describes a calibration method of using variables only observed for respondents to perform nonresponse adjustments, which may prove important in this context. Future

research needs to evaluate the ability of these alternative techniques to mitigate the effects of errors in interviewer observations on nonresponse adjustments.

This study is limited by its focus on only two interviewer observations in one large national survey in the United States. Similar in-depth investigations of both the quality of interviewer observations and the implications of poor quality for the effectiveness of alternative nonresponse adjustments are certainly needed in other survey contexts. The availability of "true" values for auxiliary variables being approximated with interviewer judgments for *nonrespondents* (in addition to respondents) would also enable study of the possible attenuating effects of errors in the judgments on response propensity models. The acquisition of administrative records containing "gold-standard" information on variables that interviewers are requested to observe (e.g., the income bracket of a household, available from tax records) for all cases in a given sample may enable this type of study in the future.

There are many avenues for future research in this area. Continuous monitoring of the quality of interviewer judgments will enable studies of trends in judgment accuracy, to see if interviewers improve as a function of experience. Such improvements would have implications for interviewer training.

Multilevel models could be used to identify respondent- and interviewer-level covariates influencing the accuracy of interviewer judgments (e.g., McCulloch et al., 2010). Survey managers could use these results to identify particular areas where observations may be

too difficult, particular interviewers who tend to produce observations of extremely high or low quality, or combinations of respondent- and interviewer-level features that lead to reduced quality in the observations.

The present study also found that interviewer estimates of the probability that a main interview will be completed are strongly predictive of response. Future work should examine the utility of these additional judgments (including their associations with key variables) in more detail. For example, one might consider assigning "scores" to interviewers based on their ability to correctly predict main interviews using this estimate, in addition to their ability to make correct judgments on current sexual activity and presence of young children (e.g., a score ranging from 0 – 3 correct for each responding case). Survey managers could then use average scores as interviewer-level covariates in subsequent models of data quality, or to identify interviewers for follow-up discussions about effective observational strategies.

The NSFG is somewhat unique in that it collects a large amount of paradata on sample units during the screening process. It also operates in a responsive survey design framework (Groves and Heeringa, 2006). Some of these paradata were correlated with the two interviewer judgments (e.g., safety concerns, physical impediments, primarily residential neighborhood, single-person household, age, etc.). Kreuter and Olson (2011) consider the case where two auxiliary variables are used for nonresponse adjustments, and show how the effectiveness of the adjustments depends heavily on the directions of the relationships of the two auxiliary variables with each other and with both response

indicators and the survey variable of interest. The present study examined the associations of interviewer observations with response indicators and key NSFG variables when adjusting for many other auxiliary variables. Future simulation studies could extend the work of Kreuter and Olson (2011), considering the effects of errors in multiple (i.e., three or more) auxiliary variables on multiple types of nonresponse adjustments. For instance, this paper considers the effects of errors in interviewer observations on simple weighting class adjustments for nonresponse, but the effects of error on more recently proposed adjustment techniques, such as calibration estimation based on optimal vectors of auxiliary variables (Sarndal and Lundstrom, 2010), deserve future research focus.

Finally, future research needs to more fully consider the implications of error-prone auxiliary variables for other survey methodologies aside from adjustment for nonresponse. This includes the use of interviewer judgments or other interviewer observations that are subject to error as predictors in propensity models used for responsive survey designs (e.g., Groves and Heeringa, 2006), in stratified sampling, and in model-based imputation approaches (e.g., sequential regression imputation; Raghunathan et al., 2001) that rely on fully observed auxiliary variables for generating predictive distributions.

Chapter IV

Factors Impacting the Accuracy of Interviewer Observations in the National Survey of Family Growth (NSFG)

Summary

Existing work showing that interviewer observations are associated with both response indicators and key survey variables in a variety of surveys has suggested that these observations may be useful auxiliary variables for constructing nonresponse adjustments. Unfortunately, the observations are typically estimates and judgments made by the interviewers, making them error-prone, and previous research (Chapter 3; Biemer et al., 2011) has suggested that errors in these types of auxiliary variables will reduce the effectiveness of nonresponse adjustments. The ability to identify both respondent- and interviewer-level factors that impact the quality of interviewer observations could assist survey researchers with the development of design strategies aimed at increasing the quality of the observations. Unfortunately, no existing studies of face-to-face surveys have attempted to identify these factors. This study attempts to fill this important gap in the literature by presenting multilevel models of observation accuracy in the National Survey of Family Growth (NSFG).

Introduction

Given previous studies demonstrating that interviewer observations recorded for survey respondents and nonrespondents alike are associated with both response indicators (Blom et al., 2011; Campanelli et al., 1997, Section 4.6; Lynn, 2003; Groves and Heeringa, 2006; Durrant and Steele, 2007; Scholes et al., 2008; Maitland et al., 2009; Blom, 2009; Casas-Cordero, 2010a; D'Arrigo et al., 2010; Durrant et al., 2010b, p. 13; Chapter 3) and key survey variables (Scholes et al., 2008; Casas-Cordero, 2010a; Gonzalez and Kreuter, 2010; Kreuter et al., 2010; Chapter 3), these observations are sometimes used to compute nonresponse adjustments for survey estimates. Unfortunately, these observations are typically judgments and estimates made by interviewers, making them prone to error (Alwin, 2008, p. 151; Sturgis and Campanelli, 1998; Casas-Cordero, 2010b; Casas-Cordero and Kreuter, 2008; Drury et al., 1980; Eckman, 2011; Feldman, 1951, p. 743; Hahn et al., 1996; Mosteller, 1944; Pickering et al., 2003; Groves et al., 2007; McCulloch et al., 2010; Tipping and Sinibaldi, 2010; Chapter 3).

The various conceptualizations of total survey error (TSE) that have been published over the years (see Groves and Lyberg, 2010, for a recent review) consistently acknowledge the problem of nonresponse bias that can arise in surveys. Errors in estimation (e.g., incorrect computation of the weights used for estimates, failure of analysts to correctly account for sample design features, etc.), unfortunately, are less often acknowledged as a key part of TSE (see Deming, 1944, or Biemer, 2010, who refers to this problem as a type of data-processing error). From a TSE perspective that also considers errors in estimation, errors in interviewer observations may lead to nonresponse adjustments that introduce *more* bias in survey estimates than was present before the nonresponse

adjustments (Biemer et al., 2011; Lessler and Kalsbeek, 1992, Ch. 8, p. 190; Stefanski and Carroll, 1985; Steiner et al., 2011; Chapter 3).

While few existing studies have *measured* the accuracy and reliability of interviewer observations, no studies of surveys using face-to-face interviewing have examined factors that impact the accuracy of the observations. Identification of such factors could provide survey methodologists with the necessary empirical evidence for developing design strategies targeted at *reducing* the error in interviewer observations collected for a specific survey. Indeed, although Steiner et al. (2011) study the implications of errors in auxiliary variables for reducing selection bias in observational studies, these authors argue for better measurement of those variables capable of reducing bias via adjustment procedures. Identification of predictors of the accuracy in auxiliary variables like interviewer observations can assist researchers with this task. Theoretical models for predicting the accuracy of personality and behavioral judgments have been proposed and applied in the psychological literature (Funder, 1995), but these approaches have not yet been applied generally in the survey methodology setting.

Drawing on findings from studies examining the accuracy of human judgments in the social psychology literature, this chapter presents exploratory multilevel analyses of factors impacting the accuracy of two interviewer observations in a personal interview survey (the National Survey of Family Growth, or NSFG). The purpose of these initial exploratory analyses is to generate hypotheses for future research by survey methodologists working in this area.

Background

Predictors of Accuracy in Behavioral or Personality Judgments

Theories from the social psychology literature may provide insights to help guide survey methodologists performing these types of exploratory analyses, especially when interviewers are tasked with recording observations that are more subjective in nature. Considering doorstep judgments about behavioral or personality characteristics of individuals (e.g., the current sexual activity of screening interview respondents in the NSFG), interviewers tasked with collecting or recording these types of paradata upon contact with sampled individuals are making what are known as *thin-slice* (or brief) observations of behaviors, based on first impressions and intuitions (e.g., Ambady and Gray, 2002; Ambady et al., 1999; Winerman, 2005). Past studies have shown that the accuracy of these thin-slice judgments can be quite high (Ambady and Rosenthal, 1992; Winerman, 2005), but studies looking at predictors of the accuracy tend to be more rare.

Examining thin-slice judgments of sexual orientation, Ambady et al. (1999) showed that accuracy was a function of how dynamic nonverbal behaviors were (i.e., judgments based on silent videos had higher accuracy than those based on still photographs) and of features of the judges relevant to the judgments (gay men and lesbians were better at judging still photos and shorter videos than heterosexuals). Women have also been shown to judge sexual orientation more accurately than men (Berger et al., 1987). Other research in this area has shown that sadness has a consistent negative impact on the accuracy of a variety of thin-slice judgments (Ambady and Gray, 2002), which could have important

practical implications for interviewers (i.e., interviewers should generally not collect observations when feeling sad or depressed). In addition, past work has shown that in cognitively demanding situations (such as listing and survey interviewing), judgments based on first impressions tend to have *higher* accuracy (Patterson and Stockbridge, 1998), which provides support for the idea of asking interviewers to make quick judgments based on first impressions. Research in this area has also shown the importance of focusing on explicit, obvious, nonverbal visual cues (e.g., hairstyles for homosexual males) for increasing the confidence and accuracy in intuitive judgments of ambiguous social categories, such as sexual orientation (Patterson et al., 2001; Rule et al., 2008). These findings suggest that interviewers having features relevant to the judgments and attending to relevant nonverbal visual cues will be more accurate in their judgments.

Interviewer judgments of behavioral traits also fit into another psychological framework for social judgment known as the "zero-acquaintance paradigm" (Albright et al., 1988; Passini and Norman, 1966), where persons are asked to make judgments about characteristics of complete strangers. Studies of zero-acquaintance situations have shown that people can in fact judge personality traits of strangers accurately based on brief observations (Ambady et al., 1999). Importantly, Ambady et al. (1995) studied *predictors* of judgment accuracy in zero-acquaintance situations, and found that judgment accuracy was higher in judges who were *female* and *less sociable*, and that male judges scoring higher on tests of nonverbal sensitivity and female judges scoring higher on tests of interpersonal perception also had higher accuracy. These findings suggest that ideal interviewers for making these types of behavioral judgments on complete strangers

should be female, less sociable and high-performing on the aforementioned tests of perceptive ability. Survey methodological studies are certainly needed to test this hypothesis, but the NSFG does not test presently test interviewers on perceptive ability, which prevented testing this hypothesis in the current study.

Tversky and Kahneman (1974) provide a theoretical framework describing mechanisms that lead to errors in judgments. Particularly relevant to the problem of errors in interviewer observations is what these authors refer to as the "representativeness" heuristic, where the probability that a judge estimates that someone falls into a certain class is based on the degree to which the person is representative of the stereotype for that class. Citing several experimental studies, these authors identify five problems with this heuristic that can lead to errors in judgments: 1) insensitivity to prior probabilities of outcomes (e.g., ignoring known proportions of the population that fall into certain categories); 2) insensitivity to sample size (i.e., larger samples will be less likely to deviate from known population features); 3) misconceptions of chance (i.e., judging that a person is in category A just because the previous *n* persons were in category B); 4) insensitivity to predictability and illusions of validity (i.e., interviewers should make judgments based on reliable, independent, and relevant predictors of what is being judged); and 5) misconceptions of regression to the mean. This framework provides guidelines that could inform interviewer training and subsequent practice in the field, assuming that observations are required.

Only one survey methodological study to date has examined both respondent- and interviewer-level factors influencing the accuracy of interviewer observations (McCulloch et al., 2010), and this study considered interviewer judgments of respondent gender in a *telephone* survey. McCulloch and colleagues found that more interviewer experience resulted in *higher* error rates when guessing respondent gender on the telephone, suggesting a hypothesis that interviewers with more experience might not take the observational task as seriously as newer interviewers with less experience. These authors also found high intra-interviewer correlations in judgment errors and significant interactions between the demographic features of respondents and interviewers when predicting judgment accuracy. There is thus a clear need for similar multilevel examinations of the factors impacting the accuracy of interviewer observations collected in face-to-face surveys.

No studies to date have considered the possibility that interviewers in the field may be using different observational *strategies* when tasked with recording particular observations, or whether different observational strategies tend to be associated with different levels of accuracy in the observations. This chapter also takes a first step in filling this gap in the existing literature by presenting an exploratory qualitative analysis of the justifications provided by interviewers for their judgments of a person-level characteristic, and examining the variance in accuracy among interviewers that is explained by the different observational strategies used. Past studies on person perception have assumed that although judges are aware of their ability to pick up nonverbal cues in making accurate judgments, they are unable to articulate the cues used (Smith et al.,

1991). The exploratory analysis presented in this chapter assesses whether interviewers can provide meaningful justifications for their observations of a person-level characteristic, and whether certain justification patterns lead to higher accuracy.

Predictors of Accuracy in Judgments of Household Characteristics

Much of this literature considers predictors of accuracy for inter-personal judgments of personality and behavioral traits, and field interviewers are often also tasked with judging features of households (e.g., income, presence of children, presence of smokers) or neighborhoods (e.g., presence of trash, evidence of social disorder, etc.; see Casas-Cordero, 2010a) rather than persons. Current research on the visual cognition phenomenon known as *inattentional blindness* (Most et al., 2005), or the failure to notice objects in plain sight when focused on a demanding visual task (such as listing or screening in surveys), has suggested that the difficulty of the observational task, rather than differences in individual ability, predicts failure to notice other objects (Simons and Jensen, 2009). This theory has been supported empirically by Casas-Cordero (2010a), who found that the probability of interviewers noting neighborhood disorders was a function of selected neighborhood characteristics and not interviewer characteristics.

This theory therefore provides support for a hypothesis that judgment accuracy may suffer in areas where observations are more difficult, as opposed to being a function of interviewer ability. For instance, housing unit access problems and urban areas are hypothesized to have a negative impact on the accuracy of housing unit observations on the presence of young children, as interviewers may have a difficult time picking up

important visual cues (e.g., children's toys) without having more ready access to the housing units. This hypothesis is also supported by the findings of Pickering et al. (2003), who suggest that rural areas tend to produce more accurate interviewer observations. If additional support for this hypothesis is found in the NSFG, survey methodologists may need to consider alternatives to interviewer observations (e.g., linked auxiliary data from purchased commercial databases) in areas predicted to result in observations with reduced accuracy. Interestingly, Kennickell (2003, p. 2123) reported that interviewers working an area probability sample for the Survey of Consumer Finances (SCF) generally dedicated more effort to following cases in apartment buildings and largely Hispanic areas, while dedicating less effort to areas with higher incomes and higher proportions of people age 65 and over. This variance in levels of effort on the part of the interviewers, depending on the features of areas or housing units, could lead to variance in the quality of the observations across areas as well.

NSFG Data

Data collected during Cycle 7 of the NSFG (July 2006 – June 2010) were analyzed in this study. Screening interviews are necessary in the NSFG to determine the eligibility of individuals in randomly selected households, given that the target population is non-institutionalized U.S. males and females aged 15-44. Additional details on the design of the NSFG, which has a primary goal of collecting "nationally representative data on factors affecting birth and pregnancy rates, family formation, and the risks of HIV and other STDs," can be found elsewhere (Groves et al., 2009).

Prior to the first face-to-face contact attempt with a randomly selected household for screening purposes, female interviewers[3] were instructed to first locate the household and then estimate whether the selected household contained any children under the age of 15 (yes / no). In the data set constructed for analyzing the amount of error in these observations for this study, there were a total of 54,733 observations on the presence of young children reported by 96 interviewers. For each of these observations, completed household roster information was available to determine whether children under the age of 15 were actually present in the household (observations with missing household roster information were deleted). There was certainly a possibility of error in the household enumeration process or the process of linking the household roster information to the interviewer judgments, but for the purposes of this study, completed household rosters were assumed to be correct and care was taken to ensure no linkage errors.

Immediately after the successful completion of the full screening interview and the selection of an eligible person from a household for the main interview, interviewers were asked to estimate whether the selected person was in a sexually active relationship with an opposite-sex partner (yes / no). There were a total of 21,340 judgments of sexual activity reported by 96 interviewers for which actual survey information on sexual activity was also available from the audio computer-assisted personal interviewing (ACASI) portion of a completed main interview. This number was reduced relative to the observations on young children because the "true" value for the variable indicating the presence of young children under the age of 15 could be measured after the household roster was completed, and did not require information from the main interview.

[3] The NSFG does not employ male interviewers for data collection.

Measurement error for the self-report of sexual activity collected in the main NSFG interview was also a real possibility, but was not considered further in this study. Care was once again taken to make sure that there were no errors in linking the interviewer judgments to the actual survey data.

This study applies multilevel multinomial logistic regression modeling (see O'Muircheartaigh and Campanelli, 1999, Pickery and Loosveldt, 2002, or Durrant and Steele, 2009 for applications to problems in survey methodology) to identify interviewer- and respondent-level predictors of the accuracy of the two judgments recorded by NSFG interviewers. This approach applies the theoretical model proposed by Funder (1995), who advocated the use of several judge- (interviewer-) and target- (respondent-) level predictors of the accuracy of *personality* judgments (given a reality about the target to be correctly judged). One categorical dependent variable will measure the accuracy of the judgments concerning presence of young children in the household [three possible values: correct, false positive (i.e., a household without children judged to have children), or false negative (i.e., a household with children judged to not have children)], and the second categorical dependent variable will measure the accuracy of the judgments about current sexual activity (correct, false positive, or false negative). This study will produce one model of accuracy for each interviewer judgment (one on a household characteristic and one on a behavioral characteristic of a respondent), and the two models will include random interviewer effects to accommodate a possible correlation of the repeated measures of accuracy for each interviewer.

The first model will examine predictors of accuracy in the judgments on the presence of young children for sampled NSFG households completing the screening interview only, given that completed household rosters are needed to determine accuracy. In total, the data set for the first model will include measures of judgment accuracy for 54,718 completed screening interviews, produced by 94 female interviewers.[4] The second model will examine predictors of accuracy only for the NSFG respondents, given that responses to the main NSFG interview are needed to determine the accuracy of judgments of sexual activity. In total, the data set for the second model includes measures of judgment accuracy for 21,340 main NSFG interviews, produced by 87 female interviewers. Table 4.1 below presents a summary of the predictors that will be considered in the models[5].

Table 4.1: Predictors at the respondent level and the interviewer level to be considered in the multilevel multinomial logistic regression models of accuracy for the two NSFG interviewer judgments.

Respondent-Level Predictors (Level 1)	Interviewer-Level Predictors (Level 2)

[4] There were 18 NSFG interviewers who chose not to complete a voluntary field researcher questionnaire (FRQ), which collected data about interviewer characteristics. These interviewers and their respondents were not included in these analyses. The 18 interviewers did not differ significantly from the 96 interviewers in terms of age, race, education, or years of experience (based on employment records). In addition, given this study's strong interest in cross-level interactions between respondent- and interviewer-level features, interviewers with less than 20 completed screening interviews (for observations on young children) and less than 20 completed main interviews (for observations on sexual activity) were dropped from the analyses, satisfying the '50/20' rule (at least 50 interviewers with at least 20 observations per interviewer) for reliable computation of random effects for all interviewers (Hox, 1998, Section 3.3).

[5] Many of the respondent-level predictors are features that were used to determine selection probabilities for persons sampled in the NSFG. The inclusion of these predictors in the models effectively accounts for the differential selection probabilities of the sampled persons, making the sampling mechanism ignorable for the purposes of constructing likelihood functions for the models and making inferences regarding the relationships of these predictors with accuracy in the target NSFG population (Little, 2008).

• Female (vs. Male)* • Age (centered within each interviewer)* • Race/Ethnicity (White, Black, Other)* • Never Married (Yes / No)* • Number of Children* • Urban Indicator (vs. Rural) for PSU • Percentage of Population in Census Zip Code Tabulation Area (ZCTA) that is Children Under 18 (Census 2000, centered within each interviewer) • Housing Unit Access Problems Observed in Segment (Yes / No) • Segment Entirely Residential (Yes / No) • Evidence of Non-English Speakers in Segment (Yes / No) • Safety Concerns in Segment (Yes / No) • Building with Many Units (Yes / No) • Physical Impediments to Housing Unit (Yes / No) • Number of Calls Made (centered within each interviewer)* • Number of Contacts Made (centered within each interviewer)* • Indicator of Any Resistance (Yes / No)* • Census Division (9 Divisions) • Ethnicity Domain of Segment (Four Domains, based on % of population that is Black and % of population that is Hispanic) • Current NSFG experience of interviewer, in days since first Cycle 7 interview (centered within each interviewer) • Indicator of measurement in Quarters 15 and 16 (when predictors of reported sexual activity were provided to interviewers)***	• Years of Interviewing Experience (centered across all interviewers) • Age (centered across all interviewers) • Race/Ethnicity (White, Black, Other) • Never Married (Yes / No) • Number of Children (centered across all interviewers) • Previous NSFG Work (Yes / No) • Interviewer Enjoyment of Doorstep Interaction (Rating from 1 to 10, with 1 being low and 10 being high; centered across all interviewers)* • College Education (Yes / No) • Other employment (Yes / No) • Variety of PSUs worked (Only Super 8, Only SR, Super 8 and SR, Super 8 and NSR, SR and NSR, All Three; Reference category = NSR only)** • Observational Strategy (for judgments of current sexual activity only)***

* Denotes predictors that are not available (or not relevant) for models of accuracy of the pre-screener housing unit observation on presence of children under the age of 15. These predictors are only available (or relevant) for models of accuracy of the interviewer observation on current sexual activity of the selected respondent, which required main interview responses for validation.

** "Super 8" PSUs are the eight largest metropolitan statistical areas (e.g., Los Angeles) appearing with certainty in each yearly quarter sample of PSUs; "Self-Representing (SR)" PSUs are the 20 next largest metropolitan statistical areas, 5 of which are rotated in each yearly quarter sample; and "Non-Self-Representing (NSR)" PSUs are the 80 randomly sampled PSUs representing smaller areas which were not selected with certainty in the NSFG sample; see Lepkowski et al. (2010) for more details.

*** These predictors were only considered in the models of accuracy for the sexual activity judgment.

Capturing the effect of providing interviewers with predictive information

The indicator variable for respondents measured in Quarters 15 and 16 will be included in the models for accuracy of the sexual activity judgment only, to capture the effect of providing interviewers with information regarding known predictors of sexual activity in these quarters to assist with their judgments (West, 2010b). This intervention was motivated by preliminary work examining significant predictors of sexual activity reports in the *main* NSFG interview that are *available to an interviewer at the time of making a judgment* (immediately after a completed screening interview), and theoretical support for improving judgment accuracy with this type of intervention also exists (Funder, 1995; Heuer, 1999; Tversky and Kahneman, 1974). The motivating hypothesis is that interviewers provided with these known predictors of reported sexual activity prior to the onset of a quarter will incorporate the factors predicting sexual activity into their judgments (making 'informed' observations that are more accurate), and preliminary analyses of data from Quarter 15 have in fact provided support for this hypothesis (West, 2010b). This hypothesis will be supported if the probability of accurate judgment for respondents measured in Quarters 15 and 16 is higher relative to earlier quarters, when controlling for other respondent-level factors (including the length of time that the interviewer had been working on NSFG when the interview was conducted).

Identifying different observational strategies used by NSFG interviewers

In these last two quarters (15 and 16) of data collection for Cycle 7 of the NSFG, the 45 interviewers were also asked to record (on laptop applications) open-ended justifications for their post-screener judgments of perceived current sexual activity for selected

persons. The interviewers were asked to provide justifications immediately after the judgments were made. However, they were not prompted for specific justifications or limited in any way. This resulted in the collection of 3,992 open-ended justifications of widely varying lengths from the 45 interviewers during these two quarters of data collection. Two examples of these justifications follow:

1. "*He works and goes to school and lives here with his twin - I do not think he could have someone over as the carpet is all taken up and it smells badly of dog poo.*" (A justification for a judgment of *not* currently sexually active.)
2. *"He has a tattoo, 'Carol', over his heart."* (A justification for a judgment of currently sexually active.)

The 3,992 justifications were coded on 13 different indicator variables (1 = mentioned in justification, 0 = not mentioned), with all indicators coded for each justification:

- Living arrangement (living with spouse, parents, alone, etc.)
- Relationship status (mention of spouse, partner, etc.)
- Age
- Household characteristics (presence of children, cultural icons, cleanliness, etc.)
- Appearance (references to physical appearance, ethnicity, or pregnancy)
- Neighborhood characteristics
- Shyness
- Guess (indication of a gut feeling, or not being sure)
- Incorrect (mention that an incorrect observation was entered in hindsight)
- Conservative (indication of conservative or strict household / parents)
- Health (reference to the health or physical disability of the person)
- Personality (reference to the person's personality or general demeanor)
- Occupation (reference to the person's occupation or student status)

In addition, the number of words used for each justification was coded as a proxy of effort dedicated to the observational task. For example, the first justification given above was coded as having 35 words, and assigned a 1 for living arrangement, household

characteristics, and occupation, and a 0 for all other indicators. All coding of the justifications was performed twice with the assistance of an undergraduate research assistant[6]. Discrepancies in coding were detected using the COMPARE procedure in the SAS software, and any discrepancies in coding or word counts were discussed and resolved. The percentage of justifications falling into each of these thirteen categories and the mean word count were then computed for each interviewer. Descriptive statistics for the interviewer-specific percentages and mean word counts are shown in Table 4.2.

Table 4.2: Descriptive statistics for interviewer justification tendencies (in descending order by mean percentages of justifications) and mean word counts.

	Mean	**SD**	**Minimum**	**Maximum**
Percentage of Justifications Mentioning:				
Relationship Status	43.25	13.70	17.86	75.00
Age	32.67	23.21	0.00	88.24
Living Arrangement	24.53	19.27	0.00	88.76
Household Characteristics	23.75	13.18	0.00	62.50
Guess	12.10	17.62	0.00	82.14
Appearance	6.86	8.22	0.00	33.00
Occupation	4.43	5.82	0.00	25.00
Personality	3.88	4.70	0.00	17.82
Health	3.32	7.89	0.00	43.14
Neighborhood Characteristics	3.16	8.77	0.00	55.41
Conservative	1.69	2.90	0.00	12.50
Incorrect	1.01	1.81	0.00	8.70
Shyness	0.39	0.91	0.00	4.00
Mean Word Count	6.32	4.03	1.90	27.92

NOTE: $n = 45$ interviewers.

A large amount of variability among the 45 interviewers is evident, in terms of the justification strategies and the average number of words used for the justifications (see Table 4.2). Interviewer justifications for sexual activity judgments most often referred to the perceived relationship status of selected respondents. All interviewers used this

[6] We are indebted to Ziming Liao from the University of Michigan Undergraduate Research Opportunity Program (UROP) for his contributions to this work.

justification for at least some of their observations and one of them for as many as 75% of the justifications made. Roughly 1% of the justifications (about 40 justifications) indicated that an incorrect judgment was entered in hindsight.

An exploratory cluster analysis was used to determine whether distinct groups of interviewers existed in terms of the percentages of justifications falling into each category and effort spent on the observational task. The 13 percentages and the mean word counts for the 45 interviewers were first standardized. An agglomerative hierarchical clustering approach was then applied using the SPSS software (Everitt et al., 2011, Chapter 4), using squared Euclidean distances based on the 14 standardized variables as distance measures between interviewers and Ward's (1963) minimum within-cluster variance method to define the clusters. This approach was selected for its established superiority in identifying known clusters when using continuous measures (Punj and Stewart, 1983).

The initial cluster analysis provided evidence of two interviewers that could be considered outliers (see Appendix B), with one interviewer citing neighborhood features in 55.41% of justifications (the next highest percentage being 17.12%), and another interviewer citing health reasons for 43.14% of justifications (the next highest percentage being 27.50%). After dropping these two interviewers, a second cluster analysis was performed that presented evidence of four clusters of interviewers based on scaled distances between the clusters (see Appendix B); that is, there were in fact distinct groups of interviewers in terms of justification tendencies. Descriptive statistics on the 14 variables for each cluster are shown in Table 4.3 below.

Table 4.3: Descriptive statistics for interviewer-level justification tendencies and mean word counts within four distinct clusters of interviewers.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	
Number of Interviewers	20	7	11	5	
Percentage of Justifications Mentioning:					Kruskal-Wallis χ^2 (df), p-value
Relationship Status	45.90 (11.10)	41.87 (7.04)	**47.76 (18.43)**	25.63 (5.85)	10.3 (3), $p = 0.016$
Age	23.02 (14.94)	49.96 (20.99)	21.03 (20.43)	**57.52 (12.52)**	17.2 (3), $p = 0.001$
Living Arrangement	**37.62 (18.36)**	20.96 (12.33)	11.11 (8.38)	2.06 (2.35)	26.1 (3), $p < 0.001$
Household Characteristics	**28.60 (13.25)**	25.11 (9.33)	13.04 (10.81)	26.23 (14.42)	9.0 (3), $p = 0.029$
Guess	5.95 (8.41)	3.96 (6.18)	**33.93 (22.14)**	1.55 (2.99)	17.0 (3), $p = 0.001$
Appearance	5.62 (4.01)	**20.60 (9.67)**	1.90 (3.89)	0.39 (0.54)	24.1 (3), $p < 0.001$
Occupation	5.59 (6.13)	**5.73 (4.41)**	1.33 (2.09)	0.00 (0.00)	14.7 (3), $p = 0.002$
Personality	3.85 (3.20)	**8.94 (6.33)**	1.09 (1.24)	0.27 (0.61)	17.0 (3), $p = 0.001$
Health	1.29 (1.61)	7.22 (10.32)	2.23 (4.62)	0.27 (0.61)	6.41 (3), $p = 0.093$
Neighborhood Characteristics	2.78 (4.71)	**3.81 (3.44)**	0.38 (1.08)	0.00 (0.00)	9.5 (3), $p = 0.023$
Conservative	2.68 (3.70)	1.75 (2.86)	0.59 (0.76)	0.00 (0.00)	5.94 (3), $p = 0.115$
Incorrect	1.29 (1.74)	0.81 (1.16)	0.24 (0.54)	2.29 (3.77)	4.06 (3), $p = 0.255$
Shyness	0.25 (0.55)	**0.98 (1.41)**	0.05 (0.16)	0.00 (0.00)	7.89 (3), $p = 0.048$
Mean Word Count	6.32 (2.48)	**7.01 (1.58)**	4.17 (1.31)	4.96 (2.29)	11.9 (3), $p = 0.008$
Gross Diff. Rate	0.247	0.191	0.168	0.171	
False Positive Rate	0.413	0.536	0.515	0.795	
False Negative Rate	0.196	0.087	0.070	0.006	

NOTES: Cells contain Mean (SD). Significance tests use a non-parametric independent samples Kruskal-Wallis test with all pairwise comparisons.

The results in Table 4.3, with the largest cluster means for each justification indicator boldfaced in the case of significant differences in distributions across the four clusters, suggest that the first cluster of interviewers is largely defined by a tendency to notice living arrangement and housing characteristics. The second cluster is largely defined by references to appearance and personality, and a relatively large word count. The third cluster is primarily defined by references to relationship status and guesses / gut feelings, while the fourth cluster focuses primarily on age, occasionally referring to relationship status and household characteristics but hardly anything else. Indicators for these derived

clusters will be included as interviewer-level covariates in exploratory multilevel models fitted to the accuracy outcome for the current sexual activity observation in Quarters 15 and 16 (later in this chapter).

Examinations of bivariate correlations and associations among the respondent-level and interviewer-level predictors separately indicated no potential problems with multicollinearity among the predictors at either level.

Statistical Analyses

Although existing theory identifies potential predictors of the accuracy in interviewer observations, the models in this study will primarily be constructed using an exploratory "step-up" model building strategy discussed by Raudenbush and Bryk (2002, p. 257) and West et al. (2006, Section 2.7.2) for multilevel modeling problems, along with general model building strategies recommended by Hosmer and Lemeshow (2000, Section 8.1.3) for multinomial logistic regression models. An initial *unconditional* model for one of the two dependent variables indicating judgment accuracy will include random interviewer effects in each of two generalized logit equations, treating a correct response as the baseline category of the dependent variable:

$$\begin{aligned} &\text{Level 1 } (i = \text{interviewer}, j = \text{respondent}): \\ &\log\left[\frac{P(\text{False Positive})_{ij}}{P(\text{Correct})_{ij}}\right] = \beta_{01i} \\ &\log\left[\frac{P(\text{False Negative})_{ij}}{P(\text{Correct})_{ij}}\right] = \beta_{02i} \end{aligned} \tag{4.1}$$

$$
\begin{aligned}
&\text{Level 2 } (i = \text{interviewer}): \\
&\beta_{01i} = \beta_{01} + u_{01i} \\
&\beta_{02i} = \beta_{02} + u_{02i} \\
&\begin{pmatrix} u_{01i} \\ u_{02i} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{01} & \sigma_{01,02} \\ \sigma_{01,02} & \sigma^2_{02} \end{pmatrix} \right]
\end{aligned}
\tag{4.2}
$$

To clarify this notation, β_{01i} represents the expected log-odds of a false positive relative to a correct judgment for interviewer *i*, β_{02i} represents the expected log-odds of a false negative relative to a correct judgment for interviewer *i*, and σ^2_{01} and σ^2_{02} are the variances of these random intercepts. Fitting these unconditional models will thus provide initial estimates of the variance across interviewers in the accuracy of observations, in addition to estimates of the *covariance* ($\sigma_{01,02}$) of the two log-odds (where a positive covariance would suggest varying levels of variable errors among interviewers, and a negative covariance would provide evidence of more systematic errors). For example, with a positive covariance, certain interviewers may have two large positive random effects, indicating a general tendency that is higher than expected for false positives *and* false negatives (variable errors), while other more accurate interviewers may have two large *negative* random effects, indicating variable errors that occur rarely. With a negative covariance, some interviewers will have larger-than-expected tendencies for false positives but *not* false negatives (systematic errors), and other interviewers will have larger than expected tendencies for false negatives but *not* false positives.

The next model will add all of the relevant respondent-level predictors to the two Level 1 logit equations. For one respondent-level predictor at a time, the two coefficients for that

predictor in the two logit equations will be allowed to vary randomly at Level 2 to assess whether effects of the predictor on accuracy vary across interviewers. An example follows below for respondent gender (represented by an indicator variable for females, $FEMALE_{ij}$):

$$\text{Level 1 } (i = \text{interviewer}, j = \text{respondent}):$$

$$\log\left[\frac{P(\text{False Positive})_{ij}}{P(\text{Correct})_{ij}}\right] = \beta_{01i} + \beta_{11i}FEMALE_{ij} + \ldots \tag{4.3}$$

$$\log\left[\frac{P(\text{False Negative})_{ij}}{P(\text{Correct})_{ij}}\right] = \beta_{02i} + \beta_{12i}FEMALE_{ij} + \ldots$$

$$\text{Level 2 } (i = \text{interviewer}):$$

$$\begin{aligned} \beta_{01i} &= \beta_{01} + u_{01i} \\ \beta_{11i} &= \beta_{11} + u_{11i} \\ \beta_{02i} &= \beta_{02} + u_{02i} \\ \beta_{12i} &= \beta_{12} + u_{12i} \\ &\ldots \end{aligned} \qquad \begin{pmatrix} u_{01i} \\ u_{11i} \\ u_{02i} \\ u_{12i} \end{pmatrix} \sim N_4\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, D_{4x4}\right] \tag{4.4}$$

The effects of the other Level 1 predictors will remain fixed at Level 2 while examining variance in the two coefficients for one of the predictors (to keep the dimensions of the variance-covariance matrices D for the four random effects at Level 2 reasonable). Random coefficients found to vary significantly ($p < 0.01$) across interviewers at Level 2 based on chi-square tests of the null hypotheses that the variances are equal to zero (Raudenbush and Bryk, 2002, pp. 63-64) will be retained in the models (for example, the effect of a female respondent on accuracy of the sexual activity observation may vary across interviewers). Only random coefficients that can be computed for more than 90 of the interviewers (for accuracy of the young children observations) and more then 80 of the interviewers (for accuracy of the sexual activity judgments) will be retained in the

models. These criteria were adopted so that the tests of significance will not be limited due to insufficient information; for example, an interviewer who interviews 22 respondents that all reside in an urban area cannot have a specific random coefficient computed representing the effect of urban vs. rural status, and will not contribute to a test of significance for the variance of the urban effects across interviewers. Respondent-level predictors with effects that do not vary randomly across interviewers will have their fixed effects tested for significance using both single-parameter (for each logit) and multi-parameter (for the two logits) tests, and will be dropped if they are not found to impact accuracy. Because no interviewer-level predictors are considered yet in this model-building step, this model will be labeled as *unconditional at Level 2*.

The final *conditional* model will then include *all* interviewer-level predictors from Table 4.1 in the Level 2 equations with retained random interviewer effects, in an effort to explain the variance in the random coefficients at Level 1 using interviewer-level predictors. The fixed effects of the interviewer-level predictors in these equations will be tested for significance (keeping the lower-order fixed effects of the interviewer-level predictors in the two Level 2 equations for the intercepts when testing the higher-order cross-level interactions), and dropped from a given Level 2 equation for both logits if not explaining a significant amount of variance in the random coefficient for at least one of the logits (beginning with the cross-level interactions first). For instance, if the fixed effects of interviewer-level predictor W are not found to be significant in at least one of the two Level 2 equations for the random coefficients of respondent-level predictor X, these fixed effects (representing cross-level interactions of W with X) will be dropped

from both logits. In other words, if an omnibus test for a given interaction term is not significant (across both logits), that interaction will be dropped entirely from the model, and if all interactions involving an interviewer-level predictor are dropped, the main effect of the interviewer-level predictor will then be tested.

Percentages of variance in the random coefficients retained at Level 1 that are explained by the significant fixed effects of the interviewer-level predictors will then be computed. For example, interviewer experience may (hypothetically) explain variance in the random intercepts in the two generalized logit equations, in addition to variance in the effects of respondent gender on accuracy across interviewers. The final estimates and inferences based on these models will indicate those variables at the respondent and interviewer levels that have the strongest relationships with judgment accuracy. This approach will also illuminate important interactions between observable respondent- and interviewer-level variables that influence accuracy (e.g., effects of access problems on accuracy might be moderated by interviewer experience), as suggested by Funder (1995).

All multilevel multinomial logistic regression models for this study will be fitted using restricted penalized quasi-likelihood (PQL) estimation (Raudenbush and Bryk, 2002, pp. 457-459), as implemented in Version 6.08 of the HLM software. Robust standard errors will be computed for all fixed effect parameter estimates to enable inferences about the fixed effects that are robust to possible misspecifications of the random effects structures of the models (Raudenbush and Bryk, 2002, pp. 276-278). These standard errors were used for making inferences because coefficients for selected respondent-level predictors

will remain fixed across interviewers, due to a lack of variance in the predictors for selected interviewers (see the explanation above for the second estimation step). The Institutional Review Board of the University of Michigan approved all procedures used in this study.

Remark: Interpenetration

Importantly, the design of the NSFG (and many other large area probability samples) does not allow for interpenetrated assignment of subsamples of the full NSFG sample to interviewers (interviewers are typically assigned to work in a single primary sampling unit only for cost efficiency). Judgment accuracy may therefore *implicitly* vary across interviewers before data collection even begins, given that judgments may be more or less difficult depending on the features of a particular PSU. The inclusion of several PSU-level and neighborhood-level features as predictors in the models (e.g., variety of PSUs worked by an interviewer, segment safety concerns, urban / rural PSU, etc.) represents an attempt to eliminate any sources of variance in accuracy either between or within interviewers due to factors beyond their control, given the design of the NSFG. Remaining components of variance due to interviewers after taking these factors into account provide an approximation of the variance in accuracy introduced by the interviewers. Methods for making inferences about components of variance due to interviewers in sample designs that do not result in assignment of interpenetrated subsamples to interviewers certainly warrant future research.

Results

Interviewer-Specific Measures of Quality

Figure 4.1 presents a scatter plot showing the association of the gross difference rates (GDRs) computed for the NSFG interviewers with valid GDRs for both judgments. Each point corresponds to a single interviewer, and the GDRs for both interviewer observations define the horizontal and vertical axes. This plot allows for an examination of whether the same interviewer tends to do well on both observation tasks. The sizes of the points in Figure 4.1 are proportional to[7] the number of judgments on presence of young children (i.e., a proxy of the number of screening interviews attempted by each interviewer). Figure 4.1 shows that there is not consistent evidence of interviewers doing poorly or doing well on both observations, as would be indicated by a linear association between the GDRs, and the weighted Pearson correlation of the two GDRs was only -0.042 ($p = 0.656$). Accuracy therefore tended to vary depending on the judgment and the interviewer, suggesting that the same interviewer may be using different strategies to make the two observations.

[7] The `survey` package in the R software was used for this analysis.

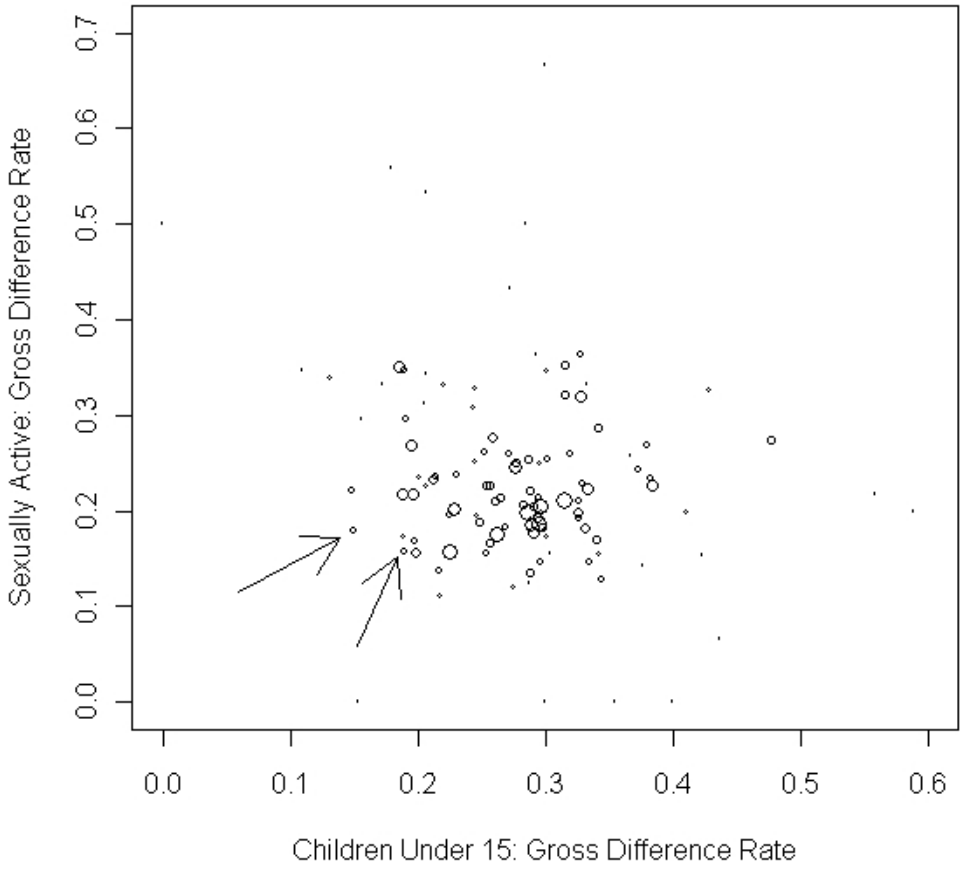


Figure 4.1: Scatter plot examining the association of interviewer gross difference rates (GDRs) on judgments of sexual activity and presence of children under age 15 in the household. Two interviewers with relatively low GDRs on both observations are highlighted with arrows. Sizes of points (weights) are based on the number of initial observations on children under 15 (representing attempted screening interviews).

Two interviewers with relatively low GDRs on both measures (indicating relatively high accuracy) are highlighted with arrows in Figure 4.1. One interviewer had 271 housing unit observations on the presence of children under 15, and was only incorrect on 51 of the observations (81.2% accuracy); this same interviewer also had 63 observations on sexual activity (after a completed screener) and was only incorrect on 11 of them (82.5% accuracy). The second interviewer had 340 housing unit observations on the presence of

children under 15, and was incorrect on 67 of them (80.3% accuracy); this same interviewer also had 113 observations on sexual activity and was only incorrect on 19 of them (83.2% accuracy). In contrast, one of the more poorly performing interviewers on both measures was incorrect on 343 out of 718 young children observations (52.2% accuracy), and 89 out of 325 sexual activity observations (72.6% accuracy).

Figure 4.1 shows evidence of interviewer variance in the accuracy of these two observations, which provides motivation for the analyses in this study. As discussed earlier, literature exists in the fields of social psychology and visual cognition that provides hypotheses for what might cause variance in accuracy between interviewers. For example, variance in accuracy between interviewers may also arise as a function of the difficulty of the primary sampling unit (PSU) being worked by an interviewer (e.g., urban areas without yards might make it harder to see children's toys). If the errors in these observations are in fact having a negative impact on subsequent nonresponse adjustments based on the judgments, then identification of the factors leading to the errors is needed.

Factors Influencing the Accuracy of Judgments about Presence of Children

Table 4.4 presents descriptive statistics for the characteristics of the 94 interviewers recording at least 20 judgments on the presence of children under the age of 15 in sampled NSFG households.

Table 4.4: Descriptive statistics for available characteristics of the 94 interviewers recording at least 20 judgments on the presence of children under the age of 15 in sampled NSFG households.

Characteristic	Mean	SD	Min	Max
Worked on NSFG Before	0.82	0.39	0	1

Age	52.43	11.39	23	87
White	0.82	0.39	0	1
Black	0.15	0.36	0	1
Other Ethnicity	0.02	0.15	0	1
Never Married	0.21	0.41	0	1
Number of Kids	1.84	1.74	0	10
College Education	0.56	0.50	0	1
Other Job	0.47	0.50	0	1
Worked Super 8 PSUs Only	0.05	0.23	0	1
Worked SR PSUs Only	0.12	0.32	0	1
Worked Super 8 and SR PSUs	0.03	0.18	0	1
Worked Super 8 and NSR PSUs	0.10	0.30	0	1
Worked SR and NSR PSUs	0.06	0.25	0	1
Worked All Three PSU Types	0.29	0.45	0	1

Table 4.4 indicates that this set of 94 interviewers is fairly heterogeneous in terms of the available characteristics, with the exceptions being high proportions of interviewers having worked on NSFG before (82%) or having been married (79%), and few interviewers of black or other ethnicity (17%). The percentages of interviewers working in various combinations of PSUs indicate a variety of experiences, with the omitted category being "worked only NSR PSUs" (35% of the 94 interviewers). These descriptive statistics provide evidence of an adequate amount of variability in the available features of the interviewers for examining the relationships of these features with the accuracy of the judgments of young children.

Figure 4.2 illustrates the amount of variance across the 94 interviewers in the percentages of judgments of presence of young children that were correct, false positives, or false

negatives. Each stacked bar in Figure 4.2 corresponds to one interviewer, with the three bars for each interviewer showing the relative percentages of judgments falling into each of the three accuracy categories.

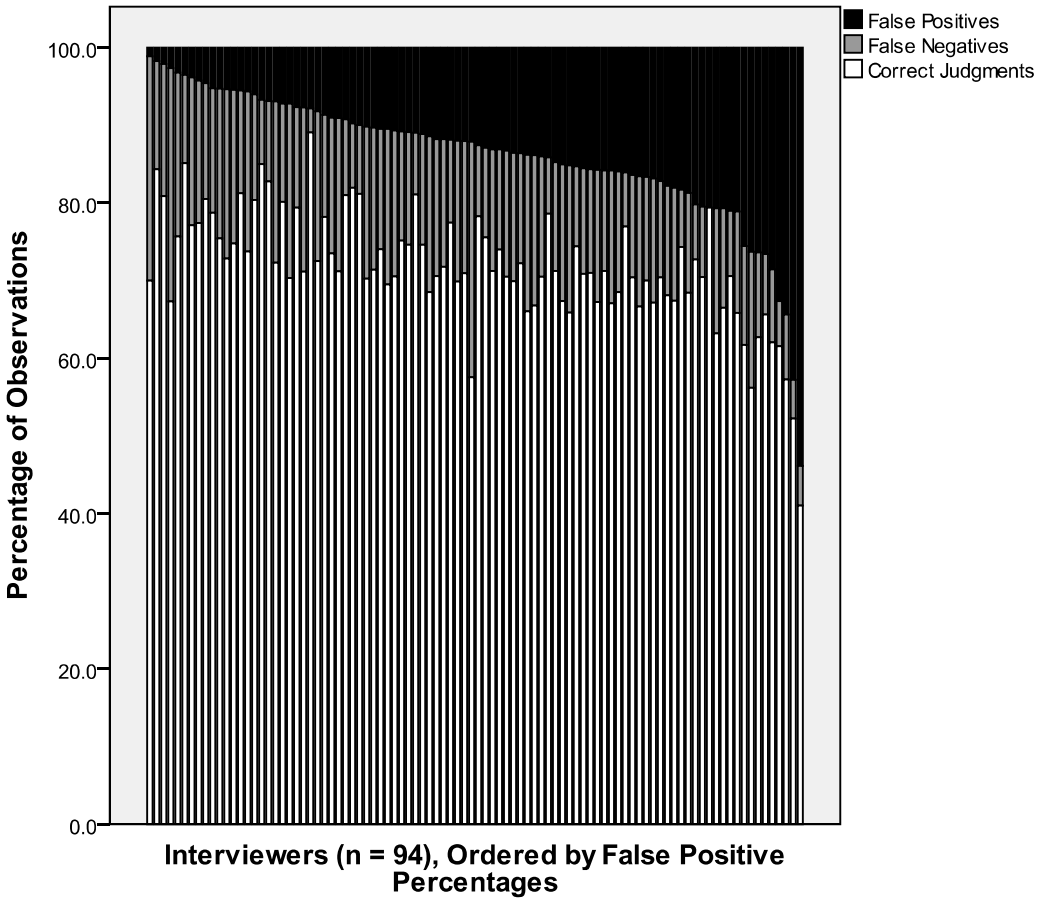


Figure 4.2: Stacked bar chart indicating variance across 94 interviewers in terms of the percentages of judgments of presence of young children that were correct, false positives, or false negatives, with interviewers ordered in terms of percentages of observations that were false positives (from lowest to highest).

As can be seen in Figure 4.2, there is a large amount of variability across interviewers in judgment accuracy, especially in terms of ratios of the percentage of false positive judgments (in black in Figure 4.2) to the percentage of correct judgments (in white in

Figure 4.2). In particular, the patterns suggest that interviewers are varying in terms of the types of systematic errors that they are making; some interviewers have a tendency to make false positive judgments and not false negative judgments, some interviewers make both types of errors, and other interviewers tend to make false negative judgments and note false positive judgments. These patterns suggest a negative covariance among interviewers in terms of the random intercepts in the initial unconditional model [as specified in (4.1) and (4.2)], and provide motivation for analyzing a three-category error indicator distinguishing between false positives and false negatives (rather than a simple binary indicator of an error vs. a correct judgment). We seek interviewer-level factors that explain portions of this variability across interviewers, in addition to respondent-level factors that may impact accuracy in different ways across interviewers. For example, the interviewer with the largest percentage of false positive judgments in Figure 4.2 (53.8% of judgments) collected all of these observations in predominantly white segments from urban areas, and this interviewer had no previous interviewing experience before starting work on the NSFG. The objective of the multilevel modeling is to assess whether patterns like this hold when considering all of the judgments recorded.

Table 4.5 presents estimates of the parameters in the final multilevel models from each estimation step for the dependent variable indicating error in the judgment of presence of young children. Results are presented in a format suggested by Raudenbush and Byrk (2002, p. 330) and Goldstein (1995, p. 107) for multilevel multinomial logistic models.

Table 4.5: Multilevel modeling results for accuracy of judgments on young children.

Fixed Effects	**Unconditional Model**	**Unconditional Model at Level 2**	**Conditional Model[a]**

For False Positive Outcome	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio
Intercept	-1.816 (0.071)	-25.465***	-2.100 (0.114)	-18.481***	-1.958 (0.137)	-14.297***
Urban PSU			0.202 (0.098)	2.059**	0.218 (0.060)	3.607***
Access Problems			-0.132 (0.070)	-1.892*	0.202 (0.079)	-2.572**
Non-English Speaking			0.150 (0.084)	1.790*	0.150 (0.039)	3.818***
Many Units			0.018 (0.064)	0.276	-0.100 (0.099)	-1.010
Physical Impediments			-0.184 (0.119)	-1.548	-0.192 (0.053)	-3.602***
Days of Experience			-0.0004 (0.0002)	-2.222**	-0.0004 (<0.0001)	-7.354***
>10% black, >10% Hisp. Segment			0.217 (0.083)	2.611**	0.213 (0.045)	4.752***
New England[b]			0.195 (0.466)	0.418	0.199 (0.153)	1.304
East North Central[b]			0.139 (0.210)	0.663	0.120 (0.084)	1.426
East South Central[b]			0.074 (0.108)	0.689	0.107 (0.117)	0.912
Mountain[b]			0.560 (0.211)	2.657***	0.561 (0.101)	5.541***
% ZCTA Under 18			0.034 (0.009)	3.697***	0.039 (0.010)	3.873***
Interviewer Black					-0.188 (0.204)	-0.922
Interviewer Never Marr.					0.096 (0.191)	0.504
Interviewer SR Only					-0.341 (0.257)	-1.328
Interviewer SUP8/SR					-0.366 (0.415)	-0.880
Interviewer SR/NSR					0.068 (0.299)	0.227
Interviewer All Three					-0.020 (0.168)	-0.119
Access Prob. x Int. NVM					0.297 (0.177)	1.673*
Many Units x Int. Black					0.492 (0.172)	2.855***
Many Units x Int. NVM					-0.511 (0.165)	-3.103***
Many Units x SUP8/SR					0.821 (0.333)	2.463**
Many Units x SR/NSR					0.642 (0.252)	2.550**
For False Negative	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio

Outcome						
Intercept	-1.628 (0.037)	-43.751***	-1.614 (0.062)	-26.130***	-1.556 (0.079)	-19.736***
Urban PSU			0.115 (0.062)	1.859*	0.116 (0.045)	2.577**
Access Problems			-0.148 (0.037)	-3.980***	-0.181 (0.037)	-4.908***
Non-English Speaking			0.110 (0.043)	2.517**	0.114 (0.033)	3.495***
Many Units			-0.107 (0.039)	-2.761***	-0.257 (0.064)	-3.992***
Physical Impediments			-0.283 (0.075)	-3.756***	-0.285 (0.047)	-6.116***
Days of Experience			0.0001 (<0.0001)	1.572	0.0001 (<0.0001)	3.009***
>10% black, >10% Hisp. Segment			0.179 (0.041)	4.377***	0.174 (0.037)	4.689***
New England[b]			-0.314 (0.126)	-2.504**	-0.322 (0.098)	-3.290***
East North Central[b]			-0.136 (0.104)	-1.310	-0.158 (0.062)	-2.564**
East South Central[b]			-0.200 (0.092)	-2.183**	-0.162 (0.089)	-1.823*
Mountain[b]			-0.158 (0.072)	-2.181**	-0.169 (0.084)	-2.019**
% ZCTA Under 18			0.035 (0.006)	6.155***	0.042 (0.006)	6.820***
Interviewer Black					0.009 (0.111)	0.085
Interviewer Never Marr.					-0.055 (0.106)	-0.522
Interviewer SR Only					-0.014 (0.145)	-0.094
Interviewer All Three					-0.020 (0.091)	-0.222
Many Units x Int. Black					-0.228 (0.110)	-2.063**
Many Units x SR Only					0.345 (0.168)	2.049**
Many Units x ALLTHR					0.337 (0.080)	4.209***
% ZCTA Under 18 x Int. NVM					-0.042 (0.015)	-2.733***
Variance Components						
	Estimate	% Var. Exp.	Estimate	% Var. Exp.	Estimate	% Var. Exp.
For False Positive Outcome						
Intercept	0.442***	--	0.408***	--	0.456***	0.00%
Access Problems			0.309***	--	0.286***	7.44%
Many Units			0.227***	--	0.178***	21.59%

% ZCTA Under 18			0.005***	--	0.005***	0.00%
For False Negative Outcome						
Intercept	0.103***	--	0.098***	--	0.110***	0.00%
Access Problems			0.021***	--	0.017**	19.05%
Many Units			0.041**	--	0.032	21.95%
% ZCTA Under 18			0.001***		0.001***	0.00%

NOTES: Baseline Outcome Category = Correct Judgment.
[a] Only significant fixed effects (and relevant lower-order terms if necessary) are shown for the conditional model. Given the omnibus nature of the fixed effects testing, if a given predictor was found to be at least marginally significant in one logit function but not the other, the non-significant fixed effect was still retained in the other logit function, but not shown in this table.
[b] The reference Census Division is a combination of West North Central, West South Central, Middle Atlantic, South Atlantic, and Pacific (where Pacific was the original reference division prior to removal of the other four indicators).
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Unconditional Model. The estimates of the variance parameters in the unconditional model [as specified in (4.1) and (4.2)] provide evidence of significant variance between interviewers in terms of both the log-odds of a false positive (relative to a correct judgment) and the log-odds of a false negative (also relative to a correct judgment). Notably, the estimated correlation of -0.367 between the random interviewer intercepts in the unconditional model (not shown in Table 4.5) suggests that interviewers tend to have high false positive rates and low false negative rates, or vice versa. This suggests that interviewer-specific errors in the judgment of young children tend to be more systematic (i.e., a given interviewer consistently makes either false positive judgments or false negative judgments, but not both). These estimates echo the illustration in Figure 4.2, where different interviewers are shown to make different types of systematic errors, and the percentages of false positives and false negatives are rarely equal for a given interviewer. These results also provide further empirical support for the multinomial

modeling approach, considering false positives and false negatives as distinct types of errors.

Unconditional Model at Level 2. Next, considering the second model that included all relevant respondent-level predictors of the accuracy of this judgment and was unconditional at the interviewer level [as specified in (4.3) and (4.4)], a multi-parameter (16 degree-of-freedom) Wald chi-square test of the null hypothesis that the fixed effects for the respondent-level predictors residential area, safety concerns, segment domain 2 (> 10% black), segment domain 3 (> 10% Hispanic), Middle Atlantic (MA) Census Division, West North Central (WNC) Census Division, South Atlantic (SA) Census Division, and West South Central (WSC) Census Division were all equal to zero in the two logit functions was not significant [$\chi^2_{16} = 19.71$, $p = 0.23$]. Many of these predictors did present evidence of significant variance in their coefficients across interviewers (e.g., residential area), but only for a subset of interviewers (due to a lack of within-interviewer variance in the predictors). As a result, these predictors were dropped from each logit function, and their estimated relationships with each logit are not shown in Table 4.5 (results of exploratory analyses examining interactions of interviewer-level features with these predictors are available in Appendix A). There was insufficient variance within at least four interviewers for the respondent-level predictors urban area (only 35 interviewers with sufficient variance), residential area (only 90 interviewers), evidence of other languages (only 78 interviewers), safety concerns (only 89 interviewers), physical impediments (only 88 interviewers), days of experience (only 64 interviewers), the three domain indicators (only 69 interviewers for the largely black domain indicator, only 61

interviewers for the largely Hispanic domain indicator, and only 69 interviewers for the black/Hispanic domain indicator), and the eight Census division indicators. In particular, the urban area indicator and Census division indicators did not vary within the majority of the interviewers, indicating that these respondent-level predictors were essentially interviewer-level features. This was not surprising, given the NSFG design of assigning most interviewers to a single primary sampling unit. Variance in random coefficients for these predictors was not tested further as a result.

First considering the odds of a false positive judgment relative to a correct judgment, households in urban areas, households in segments with evidence of languages other than English being spoken, households in segments with relatively large percentages of black and Hispanic persons, households in the Mountain Census division, and households in zip code tabulation areas (ZCTAs) with higher percentages of children under 18 all had significantly increased odds of having a false positive judgment relative to a correct judgment. This increased probability of a false positive error in areas that are more urban is fully consistent with visual cognition theory, where the difficulty of the observational task is expected to impact judgment accuracy. Interestingly, false positive judgments on young children were more likely in areas with higher percentages of young children when controlling for the other factors, suggesting that incorrect estimates of children being present may be influenced by the number of children in the area. This is consistent with the first problem with the "representativeness" heuristic proposed by Tversky and Kahneman (1974), where interviewers may simply guess that children are present

because they perceive larger numbers of children in the area, without thinking about the actual proportions of households with children.

Segments observed to have access problems (e.g., gated communities), housing units observed to have physical impediments to entry, and more experience on the part of the interviewer (in terms of days since onset of data collection) were found to result in significantly *decreased* odds of a false positive judgment relative to a correct judgment. Data collection experience thus appears to help in reducing the odds of false positives for this observation, controlling for the other factors. Interestingly, observed access problems also significantly decreased the odds of a false positive relative to a correct response, suggesting that judgments of young children tended to be more accurate in segments where access problems were observed (possibly due to the interviewers tying particular access problems to a realistic judgment of whether children would be present; e.g., young families are likely to not live in gated communities).

Next considering the odds of a false negative judgment relative to a correct judgment, households in urban areas, households in segments with evidence of languages other than English, households in segments with higher percentages of black and Hispanic persons, and households in segments with higher percentages of children under 18 had significantly higher odds of an error in judgment, only this time in the direction of a false negative. Urban areas were once again found to have households where probabilities of erroneous judgments were higher. The percentage of the population that was children was once again found to negatively impact accuracy, only making false *negative* judgments

more likely than correct judgments as well; this suggests that errors of both types were more likely in areas with higher percentages of children, and that interviewer judgments should not be based on similar area features. Interestingly, access problems, buildings with many units, physical impediments, and Census divisions aside from the Pacific, the Atlantic and the West Central all lead to significantly *reduced* odds of a false negative relative to a correct judgment on this observation, controlling for the other factors. These findings once again suggest that these household-level features simplified judgments regarding the presence of young children.

Next considering the estimated variances of (and correlations between) the eight random interviewer effects retained in this model, the relationships of segment access problems, many unit buildings, and percentage of children in the zip code under 18 with judgment accuracy were all found to vary significantly across interviewers. Therefore, in addition to evidence of significant overall effects of these predictors on accuracy, the effects of these predictors were in fact found to vary among interviewers; this finding suggests that interviewer-level features may moderate the effects of these predictors on accuracy. Two estimated correlations between random interviewer effects in the two logit functions were found to be greater than 0.4 in absolute value. Interviewers with larger random intercepts in the false positive logit (indicating higher than expected tendencies to make false positive judgments) tended to have smaller effects of the percentage of population that was young children in the false negative logit (estimated correlation = -0.478), suggesting that interviewers more likely to make false positive judgments in general were less likely to make a false negative judgment when encountering areas with more children (and vice

versa). Interestingly, the estimated correlation between the random effects of the percentage of population that was young children in the two logits was positive (0.467), suggesting that the number of young children in the area tended to impact both error probabilities in the same way for a given interviewer (i.e., for a given interviewer, higher percentages of children either led to more false positives *and* false negatives, or fewer instances of each type of error than expected). Given the significant positive relationships of the percentage of the population that is young children with the probabilities of both a false positive and a false negative, identification of (and subsequent discussions with) interviewers with *negative* random effects associated with this area-level predictor (resulting in coefficients closer to zero) may be fruitful for attempting to understand why the number of young children in the area is not decreasing observation accuracy for these particular interviewers.

Conditional Model at Level 2. In the next model building step, all of the interviewer-level predictors were added to the Level 2 equations for the eight random coefficients (four in each logit function) retained in the model. Interestingly, none of the interviewer-level predictors were found to explain the observed variance in the intercepts for either of the two logits. Indeed, the between-interviewer variance in the intercepts was actually found to *increase* slightly after adding all of the interviewer-level predictors to the Level 2 equations for the intercepts, and this remained consistent after removing non-significant fixed effects of interviewer-level predictors. No significant ($p < 0.05$) main effects of interviewer-level predictors were found for the first or the second logit.

Selected interviewer-level predictors were found to explain portions of the between-interviewer variance in the relationships of access problems and many-unit buildings with the first logit (for the odds of a false positive judgment vs. a correct judgment). For married interviewers, access problems had a significant negative effect on the first logit, decreasing the odds of a false positive relative to a correct judgment. For never married interviewers, this effect changed marginally in a positive direction, with the net effect of access problems becoming positive. This suggests that for those interviewers never having been married, access problems posed a problem for accuracy; prior marital status was found to explain about 7.5% of the observed variance in the effects of access problems on the false positive logit. Figure 4.3 below illustrates predicted probabilities of false positives (relative to correct judgments) as a function of prior marital status and access problems. This figure shows how access problems have a negative effect on accuracy among those interviewers who have never been married.

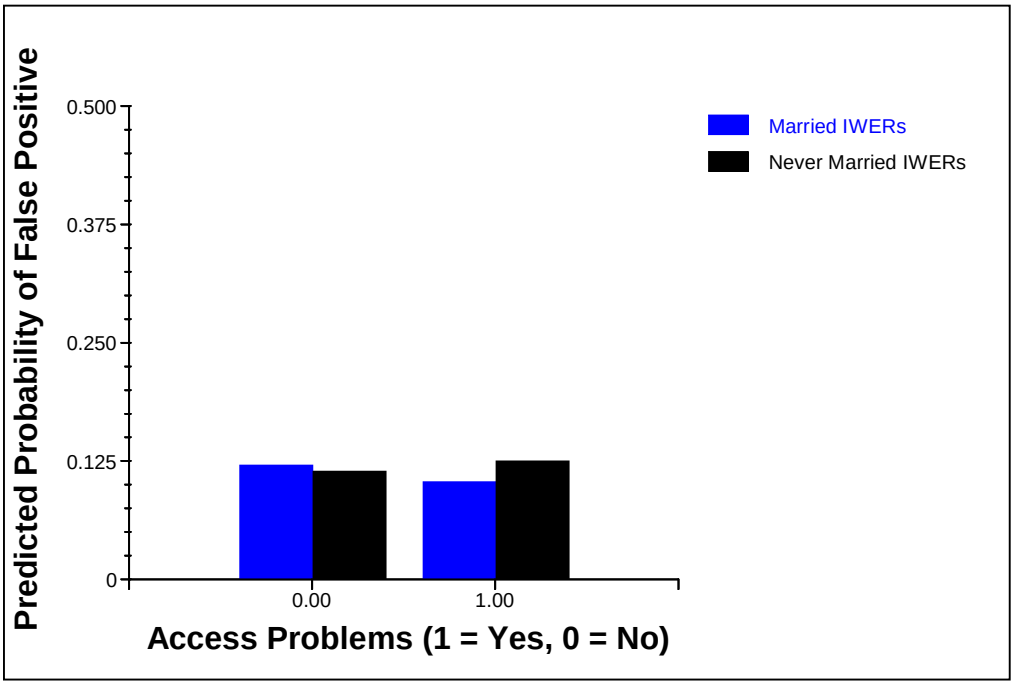


Figure 4.3: Predicted probabilities of false positives (relative to correct judgments) on presence of young children for combinations of access problems (1 = yes, 0 = no) and prior marital status, based on the final multilevel model for accuracy on this judgment (other covariates fixed to their means).

Considering white/other interviewers who were married and did not work in combinations of S8/SR or SR/NSR PSUs, many-unit buildings did not have a relationship with the odds of a false positive relative to a correct response. However, the impact of many-unit buildings on error changed in a significant positive direction for black interviewers relative to white/other interviewers, and for interviewers working in S8/SR or SR/NSR PSUs. For these interviewers, the net effect of many-unit buildings on accuracy was *positive*, indicating that false positives were more likely than correct judgments when a many-unit building was encountered. Figure 4.4 below illustrates this cross-level interaction for black interviewers relative to other interviewers, showing how encountering a many-unit building increases the probability of a false positive for black interviewers.

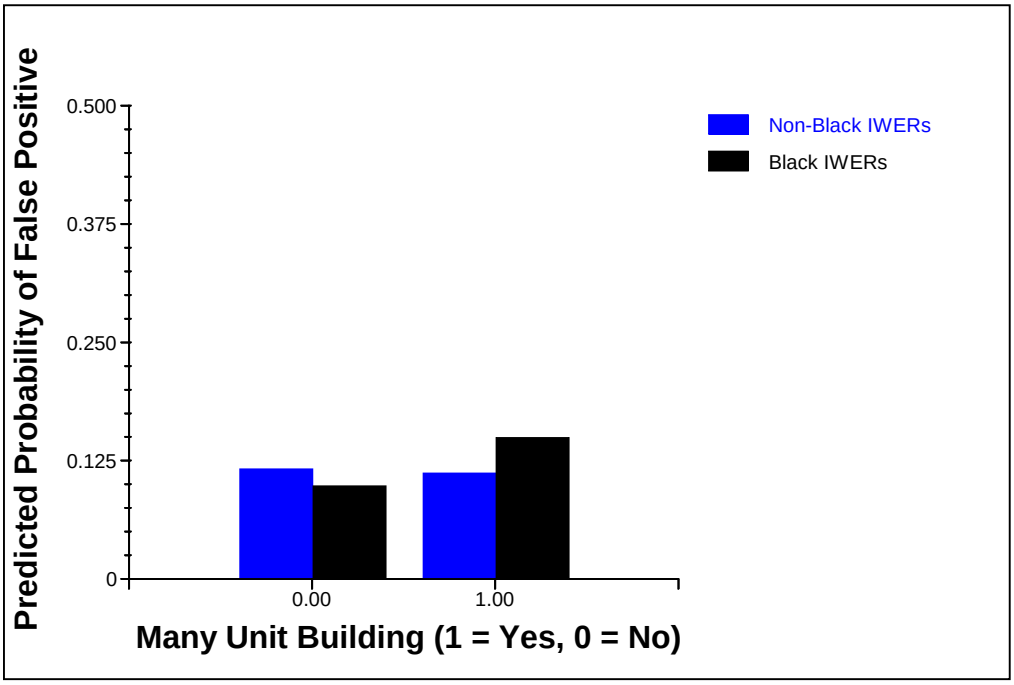


Figure 4.4: Predicted probabilities of false positives for judgments on presence of young children for combinations of many unit buildings (1 = yes, 0 = no) and Black / Non-Black interviewers, based on the final multilevel model for accuracy on this judgment (other covariates fixed to their means).

Interestingly, interviewers who had never been married had a significant *negative* change in the effect of many-unit buildings relative to the effect for married interviewers, suggesting that encountering a many-unit building reduced the odds of making a false positive (relative to a correct judgment) for never-married interviewers. Interviewers who had never been married perhaps had more experience living in many-unit buildings and were less likely to make false positive judgments, consistent with the theory of Tversky and Kahneman (1974) regarding sensitivity to prior probabilities. Figure 4.5 below illustrates this cross-level interaction. Collectively, these interviewer-level predictors were found to explain nearly 22% of the observed variance among interviewers in the

effects of many-unit buildings on the odds of a false positive relative to a correct judgment.

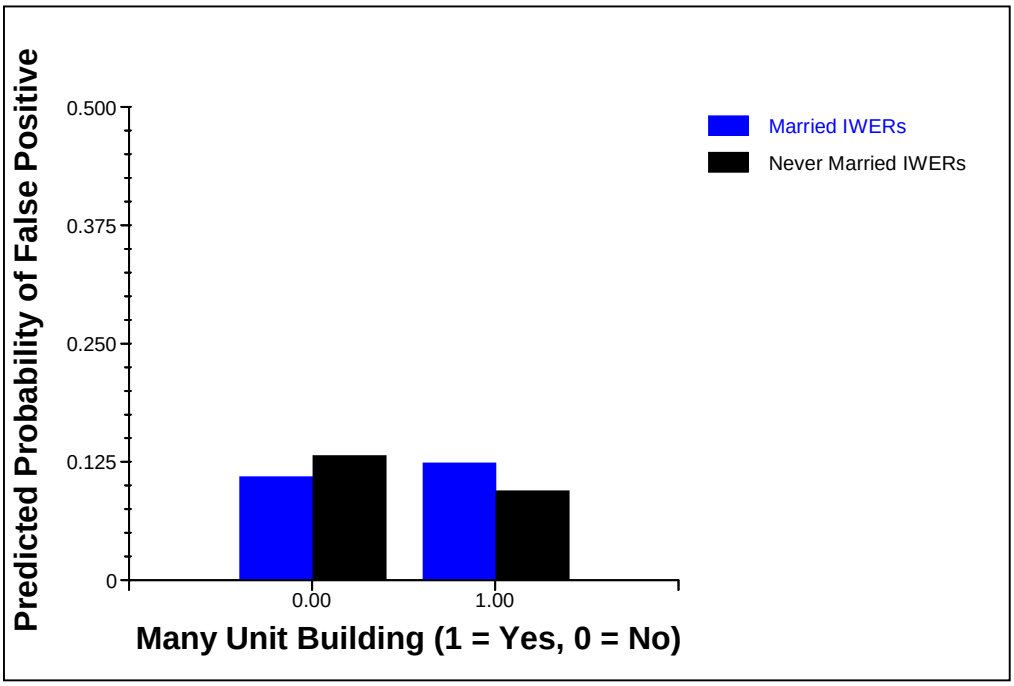


Figure 4.5: Predicted probabilities of false positives for judgments on presence of young children for combinations of many unit buildings (1 = yes, 0 = no) and never having been married for interviewers, based on the final multilevel model for accuracy on this judgment (other covariates fixed to their means).

Two interviewer-level predictors were found to explain portions of the between-interviewer variance in the relationships of many-unit buildings and percentage of the zip code area population that is children under 18 with the second logit (for the odds of a false negative judgment vs. a correct judgment). A significant negative effect of many-unit buildings was found for white/other interviewers not working in all three types of PSUs or only in SR PSUs, suggesting that encountering many-unit buildings resulted in a decreased likelihood of false negatives for these interviewers. For black interviewers, this effect became significantly more negative; therefore, when this finding is combined with

the interaction found for the false positive logit, black interviewers had significantly increased odds of a false positive and significantly decreased odds of a false negative when encountering many-unit buildings. These results suggest that black interviewers were almost exclusively making systematic false positive errors when encountering many-unit buildings. The significant negative effect of many-unit buildings essentially disappeared for interviewers working in all three PSU types, meaning that seeing buildings with many units no longer decreased the odds of a false negative (see Figure 4.6 below). This finding may have arisen due to the complexity of combinations of family structures with building types seen by interviewers working in all types of PSUs. Collectively, these predictors explained 22% of the variance across interviewers in the effects of many unit buildings on the false negative logit, and this variance component was no longer significantly greater than zero in the final conditional model.

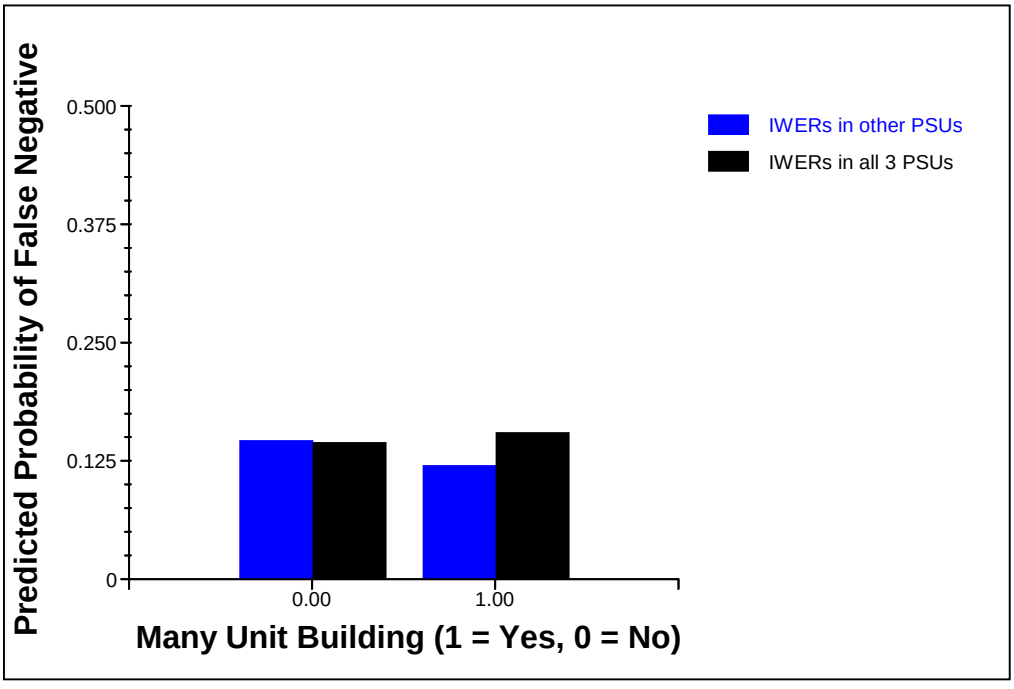


Figure 4.6: Predicted probabilities of false negatives for judgments on presence of young children for combinations of many unit buildings (1 = yes, 0 = no) and whether interviewers work all three PSU types, based on the final multilevel model for accuracy on this judgment.

The significant positive relationship of percentage of the ZCTA population that was children under 18 with the false negative logit was found to pertain to married interviewers. Interviewers who had never been married were found to have a significant *negative* change in this effect, making the net effect of this predictor on the false negative logit for never-married interviewers close to zero. Figure 4.7 below illustrates this cross-level interaction.

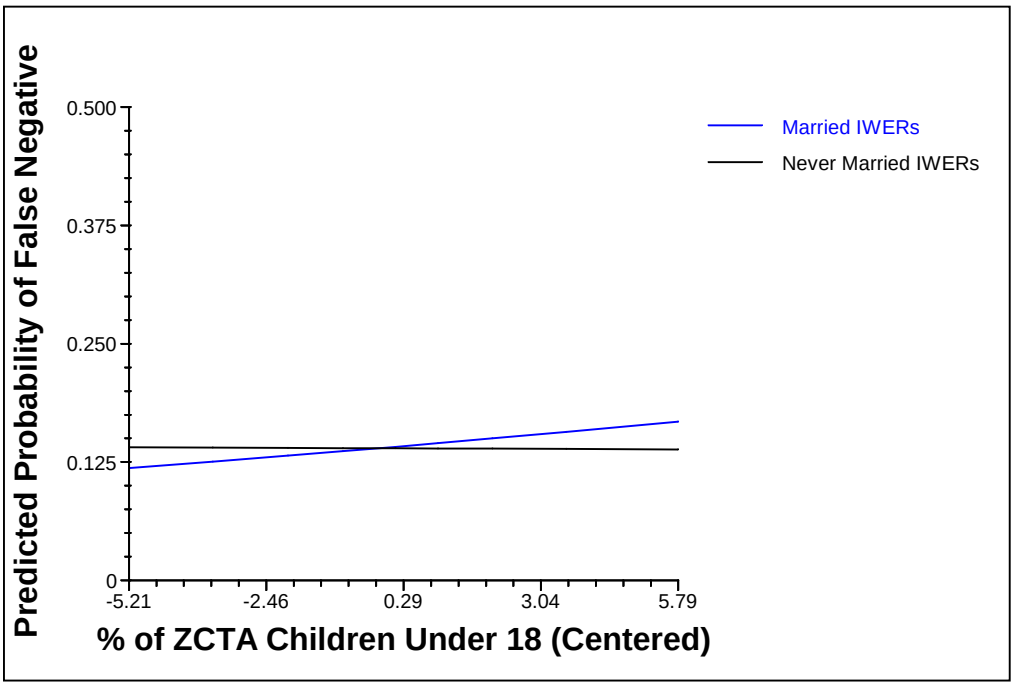


Figure 4.7: Predicted probabilities of false negatives for judgments on presence of young children as a function of percentage of ZCTA that is children under 18 (centered within interviewers) and prior marital status for interviewers, based on the final multilevel model for accuracy on this judgment (other covariates fixed to their means).

Although portions of the observed variance in the random coefficients were explained by the interviewer-level features (see Table 4.5), there was still evidence of significant unexplained variance across interviewers in seven of the eight random coefficients. This finding suggests that other interviewer-level features not studied here [e.g., observational strategies being used in the field] may be introducing the unexplained variability in the error rates and the relationships of respondent-level features with the error rates among the interviewers.

Factors Influencing the Accuracy of Judgments about Current Sexual Activity

Table 4.6 presents descriptive statistics for the characteristics of the 87 interviewers recording at least 20 judgments on the current sexual activity of eligible respondents selected from screening interviews within sampled NSFG households.

Table 4.6: Descriptive statistics for available characteristics of the 87 interviewers recording at least 20 judgments on the current sexual activity of eligible respondents selected from screening interviews within sampled NSFG households.

Characteristic	Mean	SD	Min	Max
Worked on NSFG Before	0.83	0.38	0	1
Age	53.30	10.95	27	87
White	0.82	0.39	0	1
Black	0.15	0.36	0	1
Other Ethnicity	0.02	0.15	0	1
Never Married	0.20	0.40	0	1
Number of Kids	1.90	1.76	0	10
College Education	0.57	0.50	0	1
Other Job	0.46	0.50	0	1
Worked Super 8 PSUs Only	0.05	0.21	0	1
Worked SR PSUs Only	0.09	0.29	0	1
Worked Super 8 and SR PSUs	0.03	0.18	0	1

Worked Super 8 and NSR PSUs	0.10	0.31	0	1
Worked SR and NSR PSUs	0.07	0.25	0	1
Worked All Three PSU Types	0.30	0.46	0	1

Table 4.6 indicates that this set of 87 interviewers is very similar to the set of 94 interviewers recording judgments of whether young children are present for at least 20 households (Table 4.4). This group of interviewers is also fairly heterogeneous in terms of the available characteristics, with the exceptions being high proportions of interviewers having worked on NSFG before (83%) or having been married (80%), and few interviewers of black or other ethnicity (17%). The percentages of interviewers working in various combinations of PSUs once again indicate a variety of experiences. These descriptive statistics once again provide evidence of an adequate amount of variability in the available features of the interviewers for examining the relationships of these features with the accuracy of the judgments of current sexual activity.

Figure 4.8 below illustrates the amount of variance across these 87 interviewers in the percentages of judgments of current sexual activity that were correct, false positives, or false negatives. Each stacked bar in Figure 4.8 corresponds to one interviewer, with the three bars for each interviewer showing the relative percentages of judgments falling into each of the three accuracy categories.

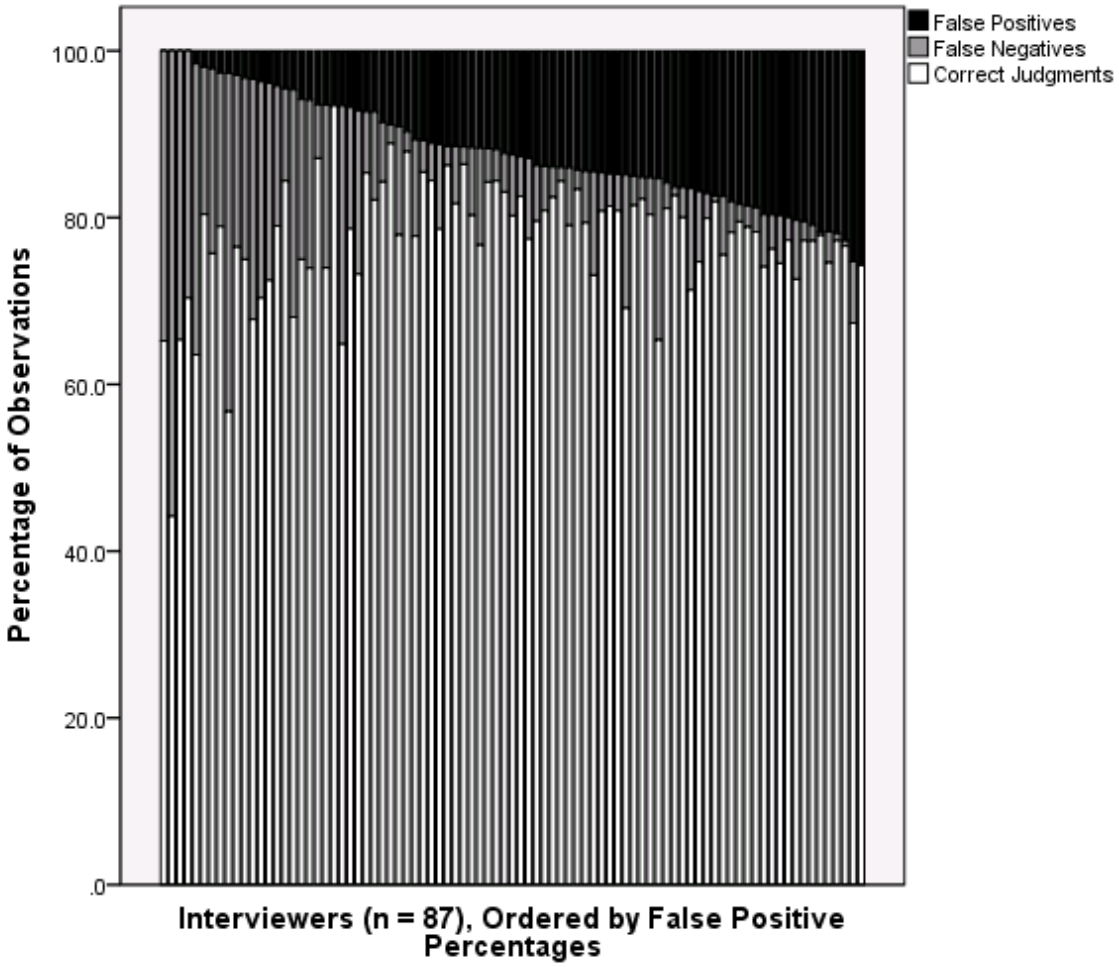


Figure 4.8: Stacked bar chart indicating variance across 87 interviewers in terms of the percentages of judgments of current sexual activity that were correct, false positives, or false negatives, with interviewers ordered in terms of percentages of observations that were false positives (from lowest to highest).

Much like the case of judgments regarding the presence of young children (Figure 4.2), there is a large amount of variability across interviewers in the accuracy of the current sexual activity judgments, especially in terms of ratios of the percentage of false negative judgments (in gray in Figure 4.8) to the percentage of correct judgments (in white in Figure 4.8). In particular, this pattern once again suggests that interviewers are varying in terms of the types of systematic errors that they are making; some interviewers have a tendency to make false positive judgments and not false negative judgments, some

interviewers make both types of errors, and other interviewers tend to make false negative judgments and not false positive judgments (with four interviewers making exclusively false negative judgments). These patterns once again suggest a negative covariance among interviewers in terms of the random intercepts in the initial unconditional model [as specified in (4.1) and (4.2)], and provide motivation for analyzing a three-category error indicator distinguishing between false positives and false negatives (rather than a simple binary indicator of an error vs. a correct judgment). We once again seek interviewer-level factors that explain portions of this variability across interviewers, in addition to respondent-level factors that may impact accuracy in different ways across interviewers.

Table 4.7 presents estimates of the parameters in the final multilevel models from each estimation step for the dependent variable indicating error in the judgment of current sexual activity.

Table 4.7: Multilevel modeling results for accuracy of judgments on current sexual activity.

Fixed Effects	**Unconditional Model**		**Unconditional Model at Level 2**		**Conditional Model[a]**	
For False Positive Outcome	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio
Intercept	-1.962 (0.061)	-32.088***	-3.190 (0.111)	-28.855***	-3.250 (0.168)	-19.339***
Intervention Quarters			-0.344 (0.091)	-3.798***	-0.353 (0.093)	-3.801***
Age of Resp.			0.014 (0.004)	3.862***	0.014 (0.004)	3.864***
Resp. Never Married			1.683 (0.083)	20.301***	1.681 (0.083)	20.316***
Resp. Female			-0.095 (0.044)	-2.151**	-0.095 (0.044)	-2.167**
Resp. No. of Kids			-0.520 (0.032)	-16.024***	-0.521 (0.032)	-16.186***

Non-English Segment			0.120 (0.056)	2.148**	0.128 (0.055)	2.316**
Safety Concerns			-0.030 (0.064)	-0.484	-0.031 (0.066)	-0.466
Many Units			-0.283 (0.056)	-5.096***	-0.285 (0.056)	-5.110***
Number of Calls			-0.015 (0.007)	-2.099**	-0.015 (0.007)	-2.101**
Days of Experience			0.0004 (0.0001)	3.457***	0.0003 (0.0001)	3.047***
Resp. Black			-0.170 (0.067)	-2.519**	-0.168 (0.068)	-2.469**
>10% Black Segment			0.009 (0.079)	0.117	0.011 (0.079)	0.138
>10% Hisp. Segment			-0.009 (0.074)	-0.115	-0.017 (0.075)	-0.227
>10% Black, >10% Hisp. Segment			0.206 (0.095)	2.178**	0.205 (0.094)	2.187**
New England[b]			-0.266 (0.281)	-0.948	-0.272 (0.274)	-0.995
Middle Atlantic[b]			-0.081 (0.113)	-0.712	-0.124 (0.110)	-1.127
East North Central[b]			-0.160 (0.124)	-1.288	-0.203 (0.124)	-1.630
West North Central[b]			-0.120 (0.117)	-1.023	-0.140 (0.110)	-1.270
East South Central[b]			-0.227 (0.256)	-0.887	-0.327 (0.258)	-1.269
West South Central[b]			-0.066 (0.116)	-0.570	-0.105 (0.109)	-0.960
Mountain[b]			-0.131 (0.115)	-1.132	-0.074 (0.107)	-0.694
Interviewer Black					0.586 (0.116)	5.051***
Interviewer Doorstep					0.071 (0.027)	2.656**
Interviewer SR Only					-0.514 (0.211)	-2.438**
Interviewer S8 / NSR					0.306 (0.183)	1.674*
For False Negative Outcome	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio	Estimate (Robust SE)	t-ratio
Intercept	-2.378 (0.112)	-21.287***	-3.010 (0.177)	-17.032***	-3.437 (0.316)	-10.883***
Intervention Quarters			0.269 (0.170)	1.581	0.273 (0.165)	1.652*
Age of Resp.			-0.001 (0.005)	-0.167	-0.001 (0.005)	-0.175
Resp. Never Married			0.802 (0.091)	8.840***	0.799 (0.091)	8.778***
Resp. Female			-0.023 (0.057)	-0.401	-0.024 (0.057)	-0.415

Resp. No. of Kids			-0.032 (0.023)	-1.350	-0.033 (0.023)	-1.391
Non-English Segment			0.024 (0.077)	0.311	0.035 (0.074)	0.473
Safety Concerns			-0.114 (0.056)	-2.039**	-0.112 (0.057)	-1.975**
Many Units			-0.004 (0.061)	-0.069	-0.010 (0.062)	-0.154
Number of Calls			0.010 (0.007)	1.410	0.010 (0.007)	1.407
Days of Experience			-0.0009 (0.0002)	-3.697***	-0.001 (<0.001)	-3.810***
Resp. Black			-0.004 (0.080)	-0.045	0.010 (0.080)	0.123
>10% Black Segment			0.235 (0.118)	1.988**	0.244 (0.121)	2.011**
>10% Hisp. Segment			0.202 (0.130)	1.553	0.198 (0.129)	1.543
>10% Black, >10% Hisp. Segment			0.195 (0.116)	1.681*	0.192 (0.116)	1.662*
New England[b]			0.678 (0.259)	2.620***	0.623 (0.256)	2.437**
Middle Atlantic[b]			0.167 (0.240)	0.695	0.190 (0.240)	0.793
East North Central[b]			-0.414 (0.187)	-2.213**	-0.335 (0.195)	-1.723*
West North Central[b]			0.090 (0.196)	0.461	0.023 (0.202)	0.112
East South Central[b]			-0.185 (0.223)	-0.828	-0.155 (0.215)	-0.723
West South Central[b]			-0.213 (0.360)	-0.592	-0.161 (0.306)	-0.526
Mountain[b]			-0.228 (0.227)	-1.005	-0.220 (0.221)	-0.996
Interviewer Prev. NSFG					0.642 (0.283)	2.271**
Interviewer Black					-1.153 (0.263)	-4.376***
Interviewer College Ed.					0.403 (0.197)	2.052**
Interviewer Doorstep					-0.074 (0.043)	-1.713*
Interviewer SR Only					0.750 (0.274)	2.733***
Interviewer S8 / NSR					-0.796 (0.379)	-2.097**
Interviewer All Three					-0.617 (0.213)	-2.894***
Variance Components						
	Estimate	% Var. Exp.	Estimate	% Var. Exp.	Estimate	% Var. Exp.
For False Positive Outcome						

Intercept	0.263***	--	0.275***	--	0.210***	23.64%
Days of Experience[c]			1.225**	--	1.296***	0.00%
For False Negative Outcome						
Intercept	0.981***	--	1.039***	--	0.723***	30.41%
Days of Experience[c]			20.736***	--	24.336***	0.00%

NOTES: Baseline Outcome Category = Correct Judgment.
[a] Only significant fixed effects (and relevant lower-order terms if necessary) are shown for the conditional model. Given the omnibus nature of the fixed effects testing, if a given predictor was found to be at least marginally significant in one logit function but not the other, the non-significant fixed effect was still retained in the other logit function, but not shown in this table.
[b] The reference Census Division is a combination of South Atlantic and Pacific (where Pacific was the original reference division prior to removal of the South Atlantic indicator).
[c] Estimates multiplied by 10^7.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Unconditional Model. Similar to the unconditional model for errors in the judgments on the presence of young children, the estimates of the variance parameters in this unconditional model [as specified in (4.1) and (4.2)] provide evidence of significant variance between interviewers in terms of both the log-odds of a false positive (relative to a correct judgment) and the log-odds of a false negative (also relative to a correct judgment). A large estimated correlation of -0.844 between the random interviewer intercepts in the unconditional model (not shown in Table 4.7) suggests that interviewers tend to have high false positive rates and low false negative rates, or vice versa. This suggests that interviewer-specific errors in the judgment of current sexual activity tend to be more systematic (i.e., a given interviewer consistently makes either false positive judgments or false negative judgments, but not both). These estimates echo the illustration in Figure 4.8, where different interviewers are shown to make different types of systematic errors, and the percentages of false positives and false negatives are rarely equal for a given interviewer. These results also provide further empirical support for the

multinomial modeling approach, considering false positives and false negatives as distinct types of errors.

Unconditional Model at Level 2. Next, considering the second model that included all relevant respondent-level predictors of the accuracy of the current sexual activity judgment and was unconditional at the interviewer level [as specified in (4.3) and (4.4)], a multi-parameter (16 degree-of-freedom) Wald chi-square test of the null hypothesis that the fixed effects for the respondent-level predictors urban PSU, access problems, physical impediments, number of contacts, any resistance, Hispanic, largely residential segment, and Census Division 5 (South Atlantic) were all equal to zero in the two logit functions was not significant [$\chi^2_{16} = 13.51$, $p = 0.369$]. These eight respondent-level predictors also did not present any evidence of variance in their coefficients across interviewers, and were thus dropped from the model. Some of the predictors with fixed effects retained in the model did present evidence of significant variance in their coefficients across interviewers (e.g., the intervention indicator), but only for a subset of interviewers (due to a lack of within-interviewer variance in the predictors). There was insufficient variance within at least eight interviewers for the following respondent-level predictors also showing evidence of interviewer variance in their coefficients: the intervention indicator (only 42 interviewers with sufficient variance, working in both quarters 1-14 and quarters 15-16), the largely black domain indicator (only 63 interviewers), and the largely Hispanic domain indicator (only 57 interviewers). Variance in the coefficients for the eight Census division indicators was not tested, based on the previous analysis of the accuracy in the young children observations. As a result, the coefficients for these

predictors remained fixed, with estimates presented in Table 4.7. Results of exploratory analyses examining interactions of interviewer-level features with these predictors are available in Appendix A.

Previous work has shown that the interviewer judgment of current sexual activity in the NSFG is prone to false positives (Chapter 3), so identification of respondent-level predictors that influence the probability of a false positive (relative to a correct judgment) could be very informative for NSFG managers. First considering the odds of a false positive judgment relative to a correct judgment, judgments in Quarters 15 and 16 [when the predictive intervention described in West (2010) was applied to all interviewers] were found to have significantly reduced odds of being false positives relative to being correct judgments when controlling for the other respondent-level predictors. The hypothesis that this intervention would be effective at increasing the quality of these sexual activity judgments was thus supported by this analysis, which is consistent with social psychological theory in this area. In addition, female respondents, black respondents, respondents with more kids, respondents in neighborhoods with safety concerns, respondents in many-unit buildings, and respondents receiving more calls all had lower odds of a false positive judgment relative to a correct judgment. The odds of false positive judgments were *increased* for older respondents, respondents who had never been married, segments with evidence of languages other than English, segments with high minority populations (>10% black, >10% Hispanic), and respondents interviewed by interviewers with increased experience in Cycle 7 [once again suggesting a reduction

in the quality of the judgments with more experience, which is consistent with results reported by McCulloch et al. (2010)].

Next, considering the odds of a false negative judgment relative to a correct judgment, there were higher odds of false negative judgments for respondents who had never been married, suggesting that errors were more likely in general for these respondents (and that interviewers may need to consider other features aside from marital status when making these judgments on individuals who have never been married). The odds of false negative judgments were also higher in segments with >10% black population and the New England Census Division (relative to the Pacific). There were *lower* odds of false negative judgments for segments with safety concerns, respondents speaking with interviewers having more NSFG experience (suggesting that more experience leads to systematic false positives), and the East North Central Census Division (relative to the Pacific).

Finally, considering the estimated variances of (and correlations between) the four random interviewer effects retained in this model, the intercepts and relationships of days of experience with judgment accuracy were found to vary significantly across interviewers in both logit functions. Therefore, although more days of experience increased the odds of false positives and decreased the odds of false negatives overall, the effects of days of experience were found to vary among interviewers. This finding suggests that interviewer-level features may moderate the effects of days of experience on accuracy, which has important training implications. Notably, the random interviewer

coefficients for the days of experience variable had a substantial negative correlation across the two logits (-0.963); interviewers with a larger than expected positive impact of days of experience on the odds of a false positive also tended to have a smaller than expected negative impact of days of experience on the odds of a false negative. The intercepts in the two logit functions were once again found to have a large negative correlation (-0.918), providing further evidence of variance in the types of systematic errors being made by interviewers on this judgment (as shown in Figure 4.10).

Conditional Model at Level 2. In the next model building step, all of the interviewer-level predictors were added to the Level 2 equations for the four random coefficients (two in each logit function) retained in the model for the accuracy of the sexual activity judgments. In contrast to the analyses of accuracy in the judgments on young children, several interviewer-level predictors were found to explain portions of the observed variance in the intercepts for the two logit functions. Considering the odds of a false positive judgment relative to a correct judgment, black interviewers, interviewers tending to like the doorstep interaction more, and interviewers working in both Super 8 and NSR PSUs were found to have significantly increased odds of a false positive, while interviewers working in SR PSUs only had significantly *decreased* odds of a false positive (relative to a correct judgment). These interviewer-level fixed effects were found to explain 23.64% of the observed variance in the intercepts of the false positive logit among the interviewers. Next considering the odds of a false negative relative to a correct judgment, interviewers with previous NSFG experience, interviewers with college education, and interviewers working in SR PSUs only were found to have significantly

increased odds of making a false negative judgment. Black interviewers, interviewers tending to like the doorstep interaction more, interviewers working in Super 8 and NSR PSUs, and interviewers working in all three types of PSUs were found to have significantly *decreased* odds of a false negative. These interviewer-level covariates explained more than 30% of the observed variance in the intercepts in the logit for false negative judgments. Collectively, these results suggest that black interviewers, interviewers liking the doorstep interaction more, and interviewers working in multiple types of PSUs tend to make systematic false positive errors on the sexual activity judgment, and NSFG managers could use this information to monitor the judgments of interviewers in these specific categories and intervene if false positives continue to be observed.

The significant negative relationship of the predictive intervention in Quarters 15 and 16 with the odds of making a false positive judgment relative to a significant judgment remained strongly significant ($p < 0.01$) in a model including all of the main effects of the interviewer-level covariates and their interactions with days of experience. The effects of other respondent-level predictors in the two logit functions also remained largely the same when including the interviewer-level covariates in the model.

Interestingly, none of the interviewer-level covariates were found to significantly moderate the relationships of days of experience with the odds of a false positive or a false negative relative to a correct judgment. These interactions were therefore dropped from the final conditional model presented in Table 4.7. The amount of unexplained

variance among interviewers in these relationships remained strongly significant when adding all interviewer-level predictors to the Level 2 equations for the random coefficients for days of experience. The variance in these relationships therefore seems to be due to other interviewer-level features that were not measured in Cycle 7 of the NSFG. In this case, one could identify interviewers with relatively extreme EBLUPs for the two random coefficients (in negative and positive directions), and inquire about the observational strategies that they are using in the field and whether these strategies are changing as data collection proceeds.

Do different observational strategies explain the variance in the accuracy of the observations on current sexual activity?

There were a total of 40 NSFG interviewers with one of four observational strategy clusters assigned and at least 20 respondents to the main NSFG interview. In total, these interviewers completed 2,846 main interviews in NSFG Quarters 15 and 16, defining the case base for the exploratory multilevel analyses designed to compare the observational strategy clusters in terms of accuracy on the current sexual activity judgments.

The initial multilevel model for accuracy of the current sexual activity judgments included all respondent-level predictors of accuracy from Table 4.1 in the two generalized logit functions, with the exception of the intervention indicator (given that all interviewers in these two quarters were exposed to the predictive intervention), the Census region indicators (given that interviewers would likely not work in multiple Census regions during these two quarters only) and the percentage of children in the zip

code under 18. Only the intercepts in the Level 1 equations were allowed to vary randomly across interviewers, given the reduced sample sizes relative to the full 16 quarter analyses presented earlier in this chapter. Future studies could consider whether observational strategies explain variance in the relationships of these predictors with observation accuracy across interviewers. The initial model also included main effects of all interviewer-level predictors from Table 4.1 in the two generalized logit equations, with the exception of indicators for working in multiple PSU types (interviewers working in multiple PSU types in these two quarters were combined with other interviewers working in NSR PSUs only, as a reference category). This was done to explain as much variance as possible in the interviewer-specific intercepts from the two logit functions prior to the addition of the indicators for the observational strategy clusters. This initial model therefore included nearly all available respondent- and interviewer-level predictors and random intercepts for the interviewers.

In the initial model, the variance of the random intercepts in the logit function for the probability of a false positive judgment relative to a correct judgment was found to be significant (estimated variance component = 0.154, Chi-square(27) = 43.866, $p = 0.021$). The variance of the random intercepts in the logit function for the probability of a false negative observation relative to a correct observation was also found to be significant (estimated variance component = 0.579, Chi-square(27) = 98.211, $p < 0.001$). These findings therefore suggested that there was a significant amount of unexplained variance in the two logit functions among interviewers even after accounting for fixed effects of

all other interviewer-level predictors, which was consistent with previous results reported in this chapter.

In the second model including fixed effects of three of the four observational strategy cluster indicators (treating cluster 4 as the reference category), the estimated variance component in the false positive logit was now 0.156, indicating that the observational strategy clusters did not explain a significant amount of the interviewer variance in the intercepts in this logit. Indeed, none of the fixed effects of the observational strategy cluster indicators were found to be significantly different from zero in the false positive logit. In contrast, the estimated variance component in the false negative logit was now 0.488, indicating that nearly 16% of the unexplained variance among interviewers in the intercepts in this logit was being accounted for by the observational strategy cluster to which an interviewer belonged. Clusters 1 and 2 were found to have marginally ($p < 0.06$) higher odds of a false negative relative to a correct judgment compared to cluster 4, suggesting that a focus of interviewers on age and household characteristics would help to reduce false negative rates on this judgment.

Discussion

This chapter has demonstrated that multilevel modeling provides survey methodologists with a useful technique for identifying respondent- and interviewer-level factors that influence the accuracy of interviewer observations. In the specific survey context of the NSFG, this study found that the accuracy of interviewer judgments of the presence of young children in households was largely a function of respondent- and area-level

features and the interactions of these features with interviewer-level characteristics, rather than interviewer-level characteristics *alone*. This finding was fully consistent with theory from the visual cognition literature which suggests that the difficulty of an observational task rather than individual ability influences one's probability of noticing a specified object. In particular, the findings suggest that interviewers are not making particular types of errors based on their own characteristics alone (e.g., how many children they have), but rather based on respondent-level features or interactions of respondent-level features with interviewer characteristics.

This study also found that the accuracy of judgments of the current sexual activity of persons selected from completed screening interviews was driven by independent effects of respondent- and interviewer-level features, rather than interactions of the features (as was found for the housing unit judgments about the presence of young children). Notably, the provision of important (and observable) predictors of current sexual activity to the interviewers at the time of an observation was found to significantly reduce the probability of a false positive for this judgment, which has been documented as a problem with this judgment in previous studies (Chapter 3). This result provides support for the hypothesis (motivated by theory from social psychology) that allowing interviewers to make this judgment when provided with observable information that is relevant to the judgment would increase the quality of the judgments. This result held when controlling for a variety of other respondent- and interviewer-level predictors, including the Cycle 7 experience of the interviewer at the time of the interview. Collectively, this study is the first to show that the accuracy of interviewer judgments of

household and respondent characteristics in a personal interview setting does in fact vary significantly as a function of respondent- and interviewer-level features, suggesting that various combinations of these features can be used to identify specific situations where interviewers struggle with the accuracy of the observations for future training purposes.

In addition, this study found evidence of distinct clusters of interviewers based on the justifications that they tended to use for their judgments of current sexual activity. This finding suggests that with only minimal guidance provided by NSFG staff, different interviewers did in fact use different observational strategies in the field when recording these judgments, and they were able to articulate reasons for their observations. This finding certainly needs to be replicated in other survey contexts to further understand this phenomenon. When controlling for other respondent- and interviewer-level characteristics, the cluster of interviewers that tended to focus on age and household characteristics during the screening interview (see Table 4.3) was found to have reduced odds of a false negative judgment (relative to a correct judgment), relative to the other clusters. The four clusters of interviewers were not found to vary in terms of the adjusted odds of a false positive judgment relative to a correct judgment.

The significant predictors of accuracy identified in the multilevel models could be used by NSFG staff for additional research and training of interviewers on methods for improving the quality of the NSFG observations in certain scenarios. The more specific findings reported in this study should be used to generate hypotheses for future research on improving the process of collecting interviewer observations. For example, given the

finding that urban areas lead to more error in judgments of the presence of young children, models predicting the presence of young children in urban areas using available paradata and auxiliary variables could be fitted and interpreted to assist interviewers with these more difficult judgments; model-based predictions might also be used *in place of* the interviewer judgments. Additional testing of whether providing interviewers with this information helps to improve accuracy will be needed in future studies. More generally, other survey organizations working in other contexts could use these techniques to identify correlates of observation accuracy, and then use this information to improve the quality of the observations with which interviewers are tasked.

In light of the relationships that these types of multilevel models can reveal, steady communication with interviewers will be essential for improving the quality of the observations. Although observation accuracy was largely found to be a function of respondent-level characteristics, it is noteworthy that the relationships of some respondent-level characteristics with accuracy were found to vary depending on interviewer-level features. For example, never-married interviewers were found to have reduced probabilities of a false positive judgment of presence of young children when encountering many-unit buildings, compared to married interviewers. Provided with this information, field supervisors could initiate conversations with interviewers from both groups to assess how observations were collected when many-unit buildings were encountered, and further understand mechanisms influencing these differences. Current research on inattentional blindness suggests that when observers adopt an *attentional set*, or ready themselves to receive a specific type of information, they are more likely to

notice other objects while performing an otherwise demanding observational task (Most et al., 2005). Given this research, conversations with interviewers who make more accurate observations when faced with certain difficulties (e.g., many-unit buildings) may reveal attentional sets (or cues to look for that indicate the presence of children in these environments) that would be useful to share with all interviewers in training sessions.

Importantly, a significant amount of unexplained variance among interviewers in both the intercepts and the effects of selected respondent-level predictors remained in the accuracy models for both observations. Multilevel modeling techniques enable the computation of predicted random effects [or Empirical Best Linear Unbiased Predictors (EBLUPs)], which in this context would allow survey managers to identify interviewers with especially unusual random coefficients. Interviewers with predicted random effects that indicate, for example, greater than expected improvements in accuracy when encountering many-unit buildings, should be contacted and asked about the observational techniques that they use in this situation; systematic collection of interviewer justifications for their observations may help to facilitate this task. Other interviewer-level features not considered in this study [e.g., features of (or attitudes about) the interviewer's own neighborhood, in the spirit of Casas-Cordero, 2010a] may also serve to explain additional variance in accuracy (or in the relationships of respondent-level features with accuracy) among interviewers.

Chapter V

Nonresponse Adjustment Based on Auxiliary Variables Subject to Error

Summary

Effective unit nonresponse adjustments require auxiliary variables that are associated with both key survey variables and the propensity to respond. Such auxiliary variables are rare in survey practice. Promising candidates include interviewer observations on sample units and linked auxiliary variables from commercially available household databases. These variables are prone to measurement error, and as a result, the assumption of missing at random (MAR) that underlies standard weighting or imputation nonresponse adjustments is violated when missingness depends on the true values of these variables, leading to biased estimates of survey quantities.

This study proposes estimators of means for this situation based on pattern-mixture models (PMMs). The set of auxiliary variables (C, X_1), measured for respondents and nonrespondents, includes one (say X_1) that is an error-prone proxy for a survey covariate (X_2). Values of X_2 and survey variables X_3 are recorded for survey respondents but missing for nonrespondents. Means of the survey variables are estimated under a PMM that assumes that nonresponse depends on C and X_2, rather than the less plausible MAR

assumption that nonresponse depends on C and X_1. Bayesian and multiple imputation estimates are developed, and compare favorably in simulation studies with imputation and weighting estimates assuming MAR, and estimates based on complete case analysis. Applications of the method to data from two real surveys are presented, together with R code for the proposed method.

Introduction

We consider nonresponse adjustments based on an auxiliary variable fully observed for a sample of n units from some population. Effective auxiliary variables for nonresponse adjustment should be highly predictive of both key survey variables and the response propensity (Beaumont, 2005; Bethlehem, 2002; Groves, 2006; Lessler and Kalsbeek, 1992; Little and Vartivarian, 2005). In an effort to collect data on auxiliary variables with these properties, several survey programs have requested that interviewers record judgments about selected features of all sample units (Kreuter et al., 2010). However, these interviewer observations are prone to measurement error (Groves et al., 2007; McCulloch et al., 2010; Pickering et al., 2003; Sturgis and Campanelli, 1998; Tipping and Sinibaldi, 2010; Chapter 3). Thus using the observations in nonresponse adjustments can be problematic. Weighting class or regression nonresponse adjustments based on error-prone auxiliary variables result in bias when missingness depends on the true underlying value (Lessler and Kalsbeek, 1992, p. 190; Chapter 3). This chapter proposes methods for correcting for this bias.

We consider data as in Figure 5.1, where X_1 is an auxiliary variable measured with error for all sampled individuals, X_2 is the underlying true value of X_1, recorded for each of r survey respondents, and X_3 is a survey variable of substantive interest, also measured for the r respondents only. The objective is to make inferences about means of the variables X_2 and X_3, using the auxiliary variable X_1 to adjust for nonresponse. The auxiliary variable X_1 may also represent a proxy variable related to key survey variables and response propensity that combines information on multiple auxiliary covariates, possibly through principal components analysis or linear predictors (e.g., Andridge and Little, 2009). As discussed later in the chapter, we can easily extend our approach to include a set of k additional auxiliary variables denoted by C that are measured without error.

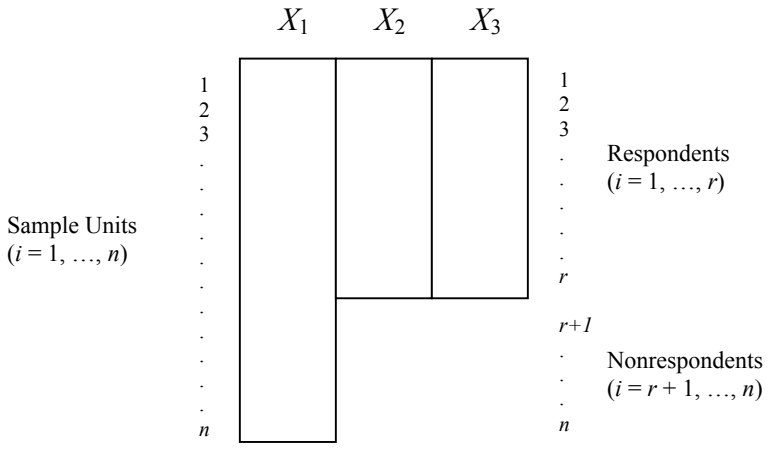


Figure 5.1: Missing data pattern under study.

Our proposed method is based on a pattern-mixture model (PMM; Little, 1994; Little and Rubin, 2002, Section 15.5), which stratifies the sample cases based on patterns of missing data and formulates distinct models for the variables within each stratum. Little (1994) derived maximum likelihood (ML) and Bayesian estimators of means and covariances for

incomplete data assuming a bivariate normal PMM, under ignorable and non-ignorable mechanisms. Little and Wang (1996) extended this work to multivariate incomplete data with fully observed covariates. More recently, Shardell et al. (2010) applied PMMs to the analysis of normal outcome data provided by proxy respondents in surveys, which may be subject to measurement error, and Baskin et al. (2011) used proxy pattern-mixture analysis, or PPMA (Andridge and Little, 2011), to estimate non-response bias in means of health expenditure variables in the Medical Expenditure Panel Survey (MEPS). In the present application, we develop a trivariate normal PMM suitable for the survey context described by Figure 5.1.

Previous studies of nonresponse adjustment based on these types of error-prone auxiliary variables have only considered ignorable missing at random (MAR) missing data mechanisms (where missingness depends on the fully observed auxiliary variable(s) only). This chapter develops PMM estimates for the case where missingness (or a failure to respond to the survey) is assumed to depend on the true auxiliary variable X_2, but not the measured variable X_1 (after conditioning on X_2), and compares their performance with complete-case estimators, weighting class estimators, and item-specific multiple imputation estimators that assume the missing data are MAR.

This work was motivated by the seventh cycle of the National Survey of Family Growth (NSFG), where interviewers were asked to judge whether sampled households had young children under the age of 15 present, and whether sampled individuals screened for participation were currently in a sexually active relationship with an opposite-sex partner

(Groves et al., 2009). Not all screened individuals participated in the survey, and analyses of NSFG data based on these interviewer judgments have suggested that the probability of responding is indeed a function of their sexual activity, with individuals judged to be sexually active having higher odds of participation (Chapter 3). At present, the NSFG uses response propensity adjustments based in part on these interviewer judgments of sexual activity to adjust base sampling weights for nonresponse (Lepkowski et al., 2010). This study aims to present an initial examination of whether item-specific PMM estimators have better empirical properties than weighting class estimators, which are primarily designed to combat unit nonresponse, in the presence of measurement error in the auxiliary variable and a non-ignorable missing data mechanism. Comparisons with multiple imputation estimates using sequential regression imputation (Raghunathan et al., 2001) are also included. Applications to real data from two different surveys are presented to illustrate use of the techniques in practice and examine potential differences in inference.

Pattern-Mixture Model: Estimation and Inference

Pattern-Mixture Model (PMM) Estimates

This section derives the PMM estimates for the context described above. For sample unit i, let m_i be a missing data indicator, equal to 0 if a unit responds to the survey request and 1 otherwise. If a unit fails to respond to the survey, it will have missing data on the variables X_2 and X_3. Then, for the missing data pattern $m_i = r$, we assume

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_1^{(r)} \\ \mu_2^{(r)} \\ \mu_3^{(r)} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^{(r)} & \sigma_{12}^{(r)} & \sigma_{13}^{(r)} \\ . & \sigma_{22}^{(r)} & \sigma_{23}^{(r)} \\ . & . & \sigma_{33}^{(r)} \end{pmatrix} \right) \equiv N_3 \left(\mu^{(r)}, \Sigma^{(r)} \right), \tag{5.1}$$

a trivariate normal distribution with nine parameters. In addition, the marginal distribution of m_i under this PMM is $m_i \sim Bernoulli(\pi_1)$. The total number of parameters in the model is 2 × 9 + 1 = 19.

The following 12 parameters are identified from the observed data in Figure 5.1: $\pi_1, \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \sigma_{11}^{(0)}, \sigma_{12}^{(0)}, \sigma_{13}^{(0)}, \sigma_{22}^{(0)}, \sigma_{23}^{(0)}, \sigma_{33}^{(0)}, \mu_1^{(1)},$ and $\sigma_{11}^{(1)}$. The following 7 parameters are not identified: $\mu_2^{(1)}, \mu_3^{(1)}, \sigma_{12}^{(1)}, \sigma_{13}^{(1)}, \sigma_{22}^{(1)}, \sigma_{23}^{(1)},$ and $\sigma_{33}^{(1)}$. Let $\beta_{\ell j.k}^{(r)}$ denote the coefficient for variable *j* in the linear regression of variable ℓ on variable *k* (*j*, *k*, and ℓ having possible values 0, 1, 2, and 3) for pattern *r*, where *j* = 0 corresponds to the intercept coefficient in the regression. Also, let $\sigma_{j\ell.k}^{(r)}$ denote the residual covariance (variance if $j = \ell$) of variable *j* and variable ℓ given variable *k* for pattern *r*. The assumption that missingness of X_2 and X_3 depends on X_2 (the "true" auxiliary variable measured in the survey) implies that the distribution of X_1 and X_3 *given* X_2 is the same for complete and incomplete cases, yielding seven parameter restrictions:

$$\beta_{10.2}^{(0)} = \beta_{10.2}^{(1)} = \beta_{10.2};\ \beta_{12.2}^{(0)} = \beta_{12.2}^{(1)} = \beta_{12.2};\ \beta_{30.2}^{(0)} = \beta_{30.2}^{(1)} = \beta_{30.2};\ \beta_{32.2}^{(0)} = \beta_{32.2}^{(1)} = \beta_{32.2};$$
$$\sigma_{11.2}^{(1)} = \sigma_{11.2}^{(0)} = \sigma_{11.2};\ \sigma_{33.2}^{(1)} = \sigma_{33.2}^{(0)} = \sigma_{33.2};\ \sigma_{13.2}^{(1)} = \sigma_{13.2}^{(0)} = \sigma_{13.2}$$

Because we have seven restrictions and seven parameters that are not identified from the observed data, the model is just-identified, and ML estimates are obtained as a straightforward extension of the methods in Little (1994).

Define the following estimates:

$(n-m)/n$ = sample proportion of nonrespondents (m = number of respondents);
$\bar{x}_j =$ sample mean of X_j for complete cases ($r = 0$) ;
$s_{jk} =$ sample covariance of X_j and X_k (variance when $j = k$) for complete cases (with denominator m)
$b_{j0.k} =$ intercept in least squares regression of X_j on X_k for complete cases ;
$b_{jk.k} =$ slope in least squares regression of X_j on X_k for complete cases ;
$s_{jj.k} =$ residual variance from least squares regression of X_j on X_k for complete cases (with denominator m)
$s_{11}^{(1)} =$ sample variance of X_1 for incomplete cases (with denominator n - m)
$\bar{x}_1^{(1)} =$ sample mean of X_1 for incomplete cases ($r = 1$)

Because the model is just-identified, the ML estimates of the 12 identified parameters are the following direct estimates:

$$\hat{\pi}_1 = (n-m)/n,\ \hat{\mu}_1^{(0)} = \bar{x}_1,\ \hat{\sigma}_{11}^{(0)} = s_{11},\ \hat{\mu}_1^{(1)} = \bar{x}_1^{(1)},\ \hat{\sigma}_{11}^{(1)} = s_{11}^{(1)}$$
$$\hat{\beta}_{10.2}^{(r)} = \hat{\beta}_{10.2} = b_{10.2}, \hat{\beta}_{12.2}^{(r)} = \hat{\beta}_{12.2} = b_{12.2}, \hat{\beta}_{30.2}^{(r)} = \hat{\beta}_{30.2} = b_{30.2}, \hat{\beta}_{32.2}^{(r)} = \hat{\beta}_{32.2} = b_{32.2},$$
$$\hat{\sigma}_{11.2}^{(r)} = \hat{\sigma}_{11.2} = s_{11.2}, \hat{\sigma}_{13.2}^{(r)} = \hat{\sigma}_{13.2} = s_{13.2}, \hat{\sigma}_{33.2}^{(r)} = \hat{\sigma}_{33.2} = s_{33.2}$$

We obtain ML estimates of the unidentified parameters by expressing them as functions of the identified parameters and the parameters defining the seven restrictions, and substituting the ML estimates. In particular, for $\mu_2^{(r)}$, we have:

$$\begin{aligned}
&\mu_1^{(r)} = \beta_{10.2}^{(r)} + \beta_{12.2}^{(r)}\mu_2^{(r)} = \beta_{10.2} + \beta_{12.2}\mu_2^{(r)} \\
&\Rightarrow \mu_2^{(r)} = \frac{\mu_1^{(r)} - \beta_{10.2}}{\beta_{12.2}} \qquad (5.2) \\
&\Rightarrow \hat{\mu}_2^{(r)} = \frac{\hat{\mu}_1^{(r)} - b_{10.2}}{b_{12.2}} = \frac{\hat{\mu}_1^{(r)} - \bar{x}_1 + b_{12.2}\bar{x}_2}{b_{12.2}} = \bar{x}_2 + \frac{\hat{\mu}_1^{(r)} - \bar{x}_1}{b_{12.2}}
\end{aligned}$$

Thus, for the marginal mean, μ_2, of X_2, we have:

$$\hat{\mu}_2 = \hat{\pi}_1\hat{\mu}_2^{(1)} + (1-\hat{\pi}_1)\hat{\mu}_2^{(0)} = \hat{\pi}_1\left[\frac{\hat{\mu}_1^{(1)} - b_{10.2}}{b_{12.2}}\right] + (1-\hat{\pi}_1)\left[\frac{\hat{\mu}_1^{(0)} - b_{10.2}}{b_{12.2}}\right] =$$

$$= \frac{\hat{\pi}_1\hat{\mu}_1^{(1)} - \hat{\pi}_1 b_{10.2}}{b_{12.2}} + \frac{(1-\hat{\pi}_1)\hat{\mu}_1^{(0)} - (1-\hat{\pi}_1)b_{10.2}}{b_{12.2}} = \frac{\hat{\mu}_1 - b_{10.2}}{b_{12.2}} = \tag{5.3}$$

$$= \frac{\hat{\mu}_1 - (\overline{x}_1 - b_{12.2}\overline{x}_2)}{b_{12.2}} = \boxed{\overline{x}_2 + \frac{\hat{\mu}_1 - \overline{x}_1}{b_{12.2}}}$$

as in Little (1994). We apply a similar approach to the other parameters:

$$\sigma_{12}^{(r)} = \beta_{12.2}^{(r)}\sigma_{22}^{(r)} = \beta_{12.2}\sigma_{22}^{(r)} = \beta_{12.2}\left[\frac{\sigma_{11}^{(r)} - \sigma_{11.2}}{\beta_{12.2}^2}\right] = \frac{\sigma_{11}^{(r)} - \sigma_{11.2}}{\beta_{12.2}}$$

$$\Rightarrow \hat{\sigma}_{12}^{(r)} = \frac{s_{11}^{(r)} - s_{11.2}}{b_{12.2}} = \frac{s_{11}^{(r)} - s_{11} + b_{12.2}^2 s_{22}}{b_{12.2}} = b_{12.2}s_{22} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}} = s_{12} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}}$$

$$\Rightarrow \hat{\sigma}_{12} = \hat{\pi}_1\hat{\sigma}_{12}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{12}^{(0)} + \hat{\pi}_1(1-\hat{\pi}_1)(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})(\hat{\mu}_2^{(1)} - \hat{\mu}_2^{(0)})$$

$$= \hat{\pi}_1\left[s_{12} + \frac{s_{11}^{(1)} - s_{11}}{b_{12.2}}\right] + (1-\hat{\pi}_1)\left[s_{12} + \frac{s_{11}^{(0)} - s_{11}}{b_{12.2}}\right] + \text{cov}, \ \{\text{note } s_{11}^{(0)} = s_{11}\}$$

$$= s_{12} + \left[\frac{\hat{\pi}_1(\hat{\sigma}_{11}^{(1)} - s_{11}) + (1-\hat{\pi}_1)(\hat{\sigma}_{11}^{(0)} - s_{11})}{b_{12.2}}\right] + \text{cov}$$

$$= s_{12} + \frac{\hat{\pi}_1\hat{\sigma}_{11}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}} + \text{cov}$$

$$\text{cov} = \hat{\pi}_1(1-\hat{\pi}_1)(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})\left[\frac{\hat{\mu}_1^{(1)} - b_{10.2}}{b_{12.2}} - \frac{\hat{\mu}_1^{(0)} - b_{10.2}}{b_{12.2}}\right] = \hat{\pi}_1(1-\hat{\pi}_1)\left[\frac{(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})^2}{b_{12.2}}\right]$$

hence

$$\boxed{\hat{\sigma}_{12} = s_{12} + \frac{\hat{\sigma}_{11} - s_{11}}{b_{12.2}}} \tag{5.4}$$

$$\sigma_{11}^{(r)} = \sigma_{11.2}^{(r)} + \beta_{12.2}^{2(r)}\sigma_{22}^{(r)} = \sigma_{11.2} + \beta_{12.2}^{2}\sigma_{22}^{(r)} \Rightarrow \sigma_{22}^{(r)} = \frac{\sigma_{11}^{(r)} - \sigma_{11.2}}{\beta_{12.2}^{2}}$$

$$\Rightarrow \hat{\sigma}_{22}^{(r)} = \frac{s_{11}^{(r)} - s_{11.2}}{b_{12.2}^{2}} = \frac{s_{11}^{(r)} - s_{11} + b_{12.2}^{2}s_{22}}{b_{12.2}^{2}} = s_{22} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}^{2}}$$

$$\Rightarrow \hat{\sigma}_{22} = \hat{\pi}_1\hat{\sigma}_{22}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{22}^{(0)} + \hat{\pi}_1(1-\hat{\pi}_1)(\hat{\mu}_2^{(1)} - \hat{\mu}_2^{(0)})^2$$

$$= \hat{\pi}_1\left[s_{22} + \frac{s_{11}^{(1)} - s_{11}}{b_{12.2}^{2}}\right] + (1-\hat{\pi}_1)\left[s_{22} + \frac{s_{11}^{(0)} - s_{11}}{b_{12.2}^{2}}\right] + \text{cov}$$

$$= s_{22} + \hat{\pi}_1\left[\frac{\hat{\sigma}_{11}^{(1)} - s_{11}}{b_{12.2}^{2}}\right] + (1-\hat{\pi}_1)\left[\frac{\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^{2}}\right] + \text{cov}$$

$$= s_{22} + \left[\frac{\hat{\pi}_1\hat{\sigma}_{11}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^{2}}\right] + \hat{\pi}_1(1-\hat{\pi}_1)\left[\frac{\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)}}{b_{12.2}}\right]^2$$

Hence $$\boxed{\hat{\sigma}_{22} = s_{22} + \frac{\hat{\sigma}_{11} - s_{11}}{b_{12.2}^{2}}} \qquad (5.5)$$

$$\mu_3^{(r)} = \beta_{30.2} + \beta_{32.2}\mu_2^{(r)} = \beta_{30.2} + \beta_{32.2}\left[\frac{\mu_1^{(r)} - \beta_{10.2}}{\beta_{12.2}}\right]$$

$$\Rightarrow \hat{\mu}_3^{(r)} = b_{30.2} + b_{32.2}\left[\frac{\hat{\mu}_1^{(r)} - b_{10.2}}{b_{12.2}}\right] = \bar{x}_3 + b_{32.2}\left[\frac{\hat{\mu}_1^{(r)} - \bar{x}_1}{b_{12.2}}\right]$$

$$\Rightarrow \hat{\mu}_3 = \hat{\pi}_1\hat{\mu}_3^{(1)} + (1-\hat{\pi}_1)\hat{\mu}_3^{(0)} =$$

$$= \hat{\pi}_1\left[b_{30.2} + b_{32.2}\left[\frac{\hat{\mu}_1^{(1)} - b_{10.2}}{b_{12.2}}\right]\right] + (1-\hat{\pi}_1)\left[b_{30.2} + b_{32.2}\left[\frac{\hat{\mu}_1^{(0)} - b_{10.2}}{b_{12.2}}\right]\right] =$$

$$= b_{30.2} + b_{32.2}\hat{\mu}_2 = b_{30.2} + b_{32.2}\left[\bar{x}_2 + \frac{\hat{\mu}_1 - \bar{x}_1}{b_{12.2}}\right]$$

Hence $$\boxed{\hat{\mu}_3 = \bar{x}_3 + b_{32.2}\left[\frac{\hat{\mu}_1 - \bar{x}_1}{b_{12.2}}\right]} \qquad (5.6)$$

$$\sigma_{13}^{(r)} = \sigma_{13.2}^{(r)} + b_{12.2}^{(r)} b_{32.2}^{(r)} \sigma_{22}^{(r)}$$

$\{$From sweep operator, conditioning on X_2 to apply restrictions$\}$

$$\Rightarrow \hat{\sigma}_{13}^{(r)} = s_{13.2} + b_{12.2} b_{32.2} \hat{\sigma}_{22}^{(r)} = s_{13.2} + b_{12.2} b_{32.2} \left[s_{22} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}^2} \right]$$

$$= s_{13} - b_{12.2} b_{32.2} s_{22} + b_{12.2} b_{32.2} s_{22} + b_{32.2} \left[\frac{s_{11}^{(r)} - s_{11}}{b_{12.2}} \right] = s_{13} + b_{32.2} \left[\frac{s_{11}^{(r)} - s_{11}}{b_{12.2}} \right]$$

$$\Rightarrow \hat{\sigma}_{13} = \hat{\pi}_1 \hat{\sigma}_{13}^{(1)} + (1 - \hat{\pi}_1) \hat{\sigma}_{13}^{(0)} + \hat{\pi}_1 (1 - \hat{\pi}_1)(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})(\hat{\mu}_3^{(1)} - \hat{\mu}_3^{(0)})$$

$$= \hat{\pi}_1 \left[s_{13} + b_{32.2} \left[\frac{s_{11}^{(1)} - s_{11}}{b_{12.2}} \right] \right] + (1 - \hat{\pi}_1) \left[s_{13} + b_{32.2} \left[\frac{s_{11}^{(0)} - s_{11}}{b_{12.2}} \right] \right] + \text{cov}$$

$$= s_{13} + b_{32.2} \left[\frac{\hat{\pi}_1 (\hat{\sigma}_{11}^{(1)} - s_{11}) + (1 - \hat{\pi}_1)(\hat{\sigma}_{11}^{(0)} - s_{11})}{b_{12.2}} \right] + \text{cov}$$

$$= s_{13} + b_{32.2} \left[\frac{\hat{\pi}_1 \hat{\sigma}_{11}^{(1)} + (1 - \hat{\pi}_1) \hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}} \right] + \text{cov}$$

$$\text{cov} = \hat{\pi}_1 (1 - \hat{\pi}_1)(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)}) \left[\bar{x}_3 + b_{32.2} \left[\frac{\hat{\mu}_1^{(1)} - \bar{x}_1}{b_{12.2}} \right] - \bar{x}_3 - b_{32.2} \left[\frac{\hat{\mu}_1^{(0)} - \bar{x}_1}{b_{12.2}} \right] \right]$$

$$= \hat{\pi}_1 (1 - \hat{\pi}_1) b_{32.2} \left[\frac{(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})}{b_{12.2}} \right]$$

$$\text{hence } \hat{\sigma}_{13} = \boxed{s_{13} + b_{32.2} \left[\frac{\hat{\sigma}_{11} - s_{11}}{b_{12.2}} \right]} \qquad (5.7)$$

$$\sigma_{23}^{(r)} = \beta_{32.2}^{(r)}\sigma_{22}^{(r)} = \beta_{32.2}\sigma_{22}^{(r)}$$

$$\Rightarrow \hat{\sigma}_{23}^{(r)} = b_{32.2}\hat{\sigma}_{22}^{(r)} = b_{32.2}\left[s_{22} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}^2}\right] = s_{23} + b_{32.2}\left[\frac{\hat{\sigma}_{11}^{(r)} - s_{11}}{b_{12.2}^2}\right]$$

$$\Rightarrow \hat{\sigma}_{23} = \hat{\pi}_1\hat{\sigma}_{23}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{23}^{(0)} + \hat{\pi}_1(1-\hat{\pi}_1)(\hat{\mu}_2^{(1)} - \hat{\mu}_2^{(0)})(\hat{\mu}_3^{(1)} - \hat{\mu}_3^{(0)})$$

$$= \hat{\pi}_1\left[s_{23} + b_{32.2}\left[\frac{\hat{\sigma}_{11}^{(1)} - s_{11}}{b_{12.2}^2}\right]\right] + (1-\hat{\pi}_1)\left[s_{23} + b_{32.2}\left[\frac{\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^2}\right]\right] + \text{cov}$$

$$= s_{23} + b_{32.2}\left[\frac{\hat{\pi}_1(\hat{\sigma}_{11}^{(1)} - s_{11}) + (1-\hat{\pi}_1)(\hat{\sigma}_{11}^{(0)} - s_{11})}{b_{12.2}^2}\right] + \text{cov}$$

$$= s_{23} + b_{32.2}\left[\frac{\hat{\pi}_1\hat{\sigma}_{11}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^2}\right] + \text{cov}$$

$$\text{cov} = \hat{\pi}_1(1-\hat{\pi}_1)\times$$

$$\left[\overline{x}_2 + \frac{\hat{\mu}_1^{(1)} - \overline{x}_1}{b_{12.2}} - \overline{x}_2 - \frac{\hat{\mu}_1^{(0)} - \overline{x}_1}{b_{12.2}}\right]\left[\overline{x}_3 + b_{32.2}\left[\frac{\hat{\mu}_1^{(1)} - \overline{x}_1}{b_{12.2}}\right] - \overline{x}_3 - b_{32.2}\left[\frac{\hat{\mu}_1^{(0)} - \overline{x}_1}{b_{12.2}}\right]\right]$$

$$= \hat{\pi}_1(1-\hat{\pi}_1)b_{32.2}\left[\frac{(\hat{\mu}_1^{(1)} - \hat{\mu}_1^{(0)})^2}{b_{12.2}^2}\right]$$

$$\text{hence } \boxed{\hat{\sigma}_{23} = s_{23} + b_{32.2}\left[\frac{\hat{\sigma}_{11} - s_{11}}{b_{12.2}^2}\right]} \qquad (5.8)$$

$$\sigma_{33}^{(r)} = \sigma_{33.2}^{(r)} + \beta_{32.2}^{2(r)}\sigma_{22}^{(r)} = \sigma_{33.2} + \beta_{32.2}^2\sigma_{22}^{(r)}$$

$$\Rightarrow \hat{\sigma}_{33}^{(r)} = s_{33.2} + b_{32.2}^2\hat{\sigma}_{22}^{(r)} = s_{33.2} + b_{32.2}^2\left[s_{22} + \frac{s_{11}^{(r)} - s_{11}}{b_{12.2}^2}\right]$$

$$= s_{33} + b_{32.2}^2\left[\frac{\hat{\sigma}_{11}^{(r)} - s_{11}}{b_{12.2}^2}\right]$$

$$\Rightarrow \hat{\sigma}_{33} = \hat{\pi}_1\hat{\sigma}_{33}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{33}^{(0)} + \hat{\pi}_1(1-\hat{\pi}_1)(\hat{\mu}_3^{(1)} - \hat{\mu}_3^{(0)})^2$$

$$= \hat{\pi}_1\left[s_{33} + b_{32.2}^2\left[\frac{\hat{\sigma}_{11}^{(1)} - s_{11}}{b_{12.2}^2}\right]\right] + (1-\hat{\pi}_1)\left[s_{33} + b_{32.2}^2\left[\frac{\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^2}\right]\right] + \text{cov}$$

$$= s_{33} + b_{32.2}^2\left[\frac{\hat{\pi}_1\hat{\sigma}_{11}^{(1)} + (1-\hat{\pi}_1)\hat{\sigma}_{11}^{(0)} - s_{11}}{b_{12.2}^2}\right] +$$

$$+\hat{\pi}_1(1-\hat{\pi}_1)\left[\overline{x}_3 + b_{32.2}\left[\frac{\hat{\mu}_1^{(1)} - \overline{x}_1}{b_{12.2}}\right] - \overline{x}_3 - b_{32.2}\left[\frac{\hat{\mu}_1^{(0)} - \overline{x}_1}{b_{12.2}}\right]\right]^2$$

$$\text{Hence } \boxed{\hat{\sigma}_{33} = s_{33} + b_{32.2}^2\left[\frac{\hat{\sigma}_{11} - s_{11}}{b_{12.2}^2}\right]} \qquad (5.9)$$

These ML estimators are unstable if the regression coefficient $b_{12.2}$ is close to zero, as when X_1 has substantial measurement error and is consequently weakly correlated with the true variable X_2. Thus, the method requires a proxy variable that has a reasonably strong correlation with the true variable.

Bayesian Inference

As indicated by Little (1994), large-sample standard errors for the ML estimates derived above can be based on linearized variance estimators. A better approach for small samples is to assume noninformative prior distributions and simulate draws from the posterior distribution of the parameters. In addition, in the case of larger measurement error in the auxiliary variable (leading to weaker associations with the survey variables) and stronger dependence of missingness on the true auxiliary variable, the coverage of large-sample confidence intervals based on ML estimates has been shown in previous simulation studies to be below nominal compared to intervals based on Bayesian approaches (Andridge and Little, 2011, p. 166).

The methods for bivariate normal incomplete data in Little (1994) can be readily extended to our case of trivariate normal incomplete data. Specifically, we assume Jeffreys' noninformative priors for the 12 identified parameters:

$$\pi_0 \sim Beta(0.5, 0.5)$$
$$p(\mu^{(0)}, \Sigma^{(0)}) \propto \left|\Sigma^{(0)}\right|^{-1}$$
$$p(\mu_1^{(1)}, \sigma_{11}^{(1)}) \propto 1/\sigma_{11}^{(1)}$$

Draws from the posterior distribution of the identified parameters are obtained as follows:

$$1)\ \pi_0^{(d)} \sim Beta(n_0 + 0.5,\ n_1 + 0.5),\ \pi_1^{(d)} = 1 - \pi_0^{(d)};\quad 2)\ \sigma_{22}^{(0)(d)} \sim \frac{n_0 s_{22}}{\chi^2_{n_0-1}};$$

$$3)\ \sigma_{11}^{(1)(d)} \sim \frac{n_1 s_{11}^{(1)}}{\chi^2_{n_1-1}};\quad 4)\ \sigma_{11.2}^{(0)(d)} \sim \frac{n_0 s_{11.2}}{\chi^2_{n_0-2}};\quad 5)\ b_{12.2}^{(d)} \sim N(b_{12.2}, \frac{\sigma_{11.2}^{(0)(d)}}{n_0 s_{22}});$$

$$6)\ b_{10.2}^{(d)} \sim N(\bar{x}_1 - b_{12.2}^{(d)}\bar{x}_2,\ \sigma_{11.2}^{(0)(d)} / n_0);\quad 7)\ \mu_2^{(0)(d)} \sim N(\bar{x}_2,\ \sigma_{22}^{(0)(d)} / n_0);$$

$$8)\ \mu_1^{(1)(d)} \sim N(\bar{x}_1^{(1)},\ \sigma_{11}^{(1)(d)} / n_1);\quad 9)\ \sigma_{33}^{(0)(d)} \sim \frac{n_0 s_{33}}{\chi^2_{n_0-1}};\quad 10)\ \sigma_{33.2}^{(0)(d)} \sim \frac{n_0 s_{33.2}}{\chi^2_{n_0-2}};$$

$$11)\ b_{32.2}^{(d)} \sim N(b_{32.2}, \frac{\sigma_{33.2}^{(0)(d)}}{n_0 s_{33}});\quad 12)\ b_{30.2}^{(d)} \sim N(\bar{x}_3 - b_{32.2}^{(d)}\bar{x}_2,\ \sigma_{33.2}^{(0)(d)} / n_0)$$

To satisfy the constraints on these parameters, the draw in 3) must be greater than the draw in 4) (Little, 1994). If this is not the case, the draws are discarded and these draws are repeated. The same condition applies to the draws in steps 9) and 10) as well. The drawn values from the sequence above then replace the ML estimates in (5.2) to (5.9) to generate draws from the posterior distributions of the other parameters. Inferences about the parameters are based on a large sample (say, 1,000) of these draws. In particular, the mean of the draws simulates the posterior mean, and the 2.5% and 97.5% percentiles of the simulated draws simulate a 95% credible interval for the mean.

Multiple Imputation

As discussed in Andridge and Little (2011), an alternative method for making inferences about means in this context is multiple imputation (Little and Rubin, 2002). This method involves the creation of B complete data sets, where missing values on X_2 and X_3 are imputed from their posterior distributions based on the PMM. Draws from the posterior distributions of X_2 and X_3 are obtained by first generating a set of draws from the posterior distributions of the parameters in the PMM as described above. The missing

values of X_2 are then drawn from the following conditional distribution of X_2 given X_1 for nonrespondents ($m_i = 1$):

$$[x_{2i} \mid x_{1i}, m_i = 1, \phi_{(b)}] \sim N\left(\mu_{2(b)}^{(1)} + \frac{\sigma_{12(b)}^{(1)}}{\sigma_{11(b)}^{(1)}}\left(x_{1i} - \mu_{1(b)}^{(1)}\right), \sigma_{22(b)}^{(1)} - \frac{\left(\sigma_{12(b)}^{(1)}\right)^2}{\sigma_{11(b)}^{(1)}} \right) \tag{5.10}$$

In this notation, (b) denotes the b-th set of draws of the PMM parameters, collectively denoted by $\phi_{(b)}$, where $b = 1, \ldots, B$. After missing values for X_2 have been imputed based on draws from (5.10), a similar predictive distribution for X_3 as a function of X_1 and the imputed X_2 (so as to maintain the association between X_2 and X_3 in the imputed data set) can be defined as follows:

$$\begin{aligned} &[x_{3i} \mid \begin{pmatrix} x_{1i} \\ x_{2i(b)} \end{pmatrix} \equiv x_i, m_i = 1, \phi_{(b)}] \\ &\sim N\left(\begin{array}{l} \mu_{3(b)}^{(1)} + \begin{pmatrix} \sigma_{31(b)}^{(1)} & \sigma_{32(b)}^{(1)} \end{pmatrix} \begin{pmatrix} \sigma_{11(b)}^{(1)} & \sigma_{12(b)}^{(1)} \\ \sigma_{12(b)}^{(1)} & \sigma_{22(b)}^{(1)} \end{pmatrix}^{-1} \left(x_i - \begin{pmatrix} \mu_{1(b)}^{(1)} \\ \mu_{2(b)}^{(1)} \end{pmatrix} \right), \\ \sigma_{33(b)}^{(1)} - \begin{pmatrix} \sigma_{31(b)}^{(1)} & \sigma_{32(b)}^{(1)} \end{pmatrix} \begin{pmatrix} \sigma_{11(b)}^{(1)} & \sigma_{12(b)}^{(1)} \\ \sigma_{12(b)}^{(1)} & \sigma_{22(b)}^{(1)} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{31(b)}^{(1)} \\ \sigma_{32(b)}^{(1)} \end{pmatrix} \end{array} \right) \end{aligned} \tag{5.11}$$

Note that $x_{2i(b)}$ denotes the imputed value for x_{2i} based on the b-th set of draws of the PMM parameters. The missing values of X_3 are imputed based on random draws from the predictive distribution in (5.11).

Given the b-th complete data set following this method, the estimates of the means for X_2 and X_3 are simply the sample means (appropriately weighted if analyzing complex sample survey data). Standard errors for the estimated means based on the b-th complete data set can then be computed based on the sample design features. This method of inference will therefore appeal to analysts of complex sample survey data, given that

complex sample design features can be accommodated in the complete case analyses (see Heeringa et al., 2010). After *B* estimates of each mean and their standard errors have been computed from the *B* complete data sets, standard combining rules for multiple imputation inference can be applied to obtain consistent estimates of the means in addition to standard errors for the estimates that incorporate within- and between-imputation variance in the estimates (Little and Rubin, 2002, Chapter 5). Importantly, this method would also allow survey organizations to use relevant (and possibly restricted) auxiliary information when imputing missing values on survey variables using PMMs, and then release multiple versions of complete data sets to the public, with guidelines for performing multiple imputation analyses of the data.

Simulation Studies

We used simulations to assess empirically the performance of the derived PMM estimates (using Bayesian methods for inference) relative to other popular methods of compensating for unit nonresponse in surveys.

Simulations Based on the Normal Selection Model

The first set of simulations considered the case when unit nonresponse arises from a nonignorable selection model. We first selected a sample of size n = 1,000 from the trivariate normal model

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 1 \\ 1 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0.25 \\ \rho & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} \right).$$

In this model, the parameter ρ determines the measurement error in X_1, and is set to 0.9 for low measurement error and 0.6 for high measurement error. The X_1 variable has a weaker association with X_3 than the "true" auxiliary variable X_2, to reflect attenuation of the relationships due to measurement error in X_1 (Fuller, 1987).

To create missing data, values of X_2 and X_3 were deleted using the model

$$P(m_i = 0 \mid x_{i2}, \alpha, \beta) = \frac{\exp(\alpha + \beta x_{i2})}{1 + \exp(\alpha + \beta x_{i2})},$$

where α (with possible values 0 and -1) determines the expected response rate, and β (with possible values 2, 1, and 0) determines the dependence of response on the "true" auxiliary variable X_2, allowing for analyses of sensitivity to assumptions about the non-ignorable missing data mechanism. For each sample case, a random UNIFORM(0,1) deviate was drawn, and the values of X_2 and X_3 were retained if this draw was less than or equal to $P(m_i = 0 \mid x_{i2}, \alpha, \beta)$, and deleted otherwise.

Four approaches to estimation and inference for the means of the variables X_2 and X_3 were then applied to each of the 1,000 samples:

1) PMM estimates and 95% credible intervals for the means based on the Bayesian approach described above. Given that the data are not simulated from a pattern-mixture model in this case, we can assess whether the PMM method provides inferences that are robust to this particular form of model misspecification.

2) A multiple imputation (MI) approach assuming an ignorable missing data mechanism, where missing values on X_2 and X_3 are imputed multiple (5) times using an iterative conditional sequential regression imputation approach, as implemented in the `mi` package of R (Su et al., 2009). Missing values on X_2 are imputed using conditional draws from the predictive distribution for a linear regression model of X_2 on X_1, and missing values on X_3 are imputed using conditional draws from the predictive distribution for a linear regression model of X_3 on the imputed and observed values of X_2 and the fully observed values of X_1. Combining rules described by Rubin (1987) and discussed further in Little and Rubin (2002) are used for computing multiple imputation estimates of the two means and standard errors of the estimates based on the five imputed data sets. Degrees of freedom for computing 95% confidence intervals for the means are computed using the methods for large samples described by Barnard and Rubin (1999, Section 1).

3) A "global" weighting (GW) approach that also assumes an ignorable missing data mechanism, where the probability of response for each case was estimated based on a logistic regression model with a *response* indicator ($1 - m_i$) as a dependent variable, and the fully observed error-prone auxiliary variable X_1 as an independent variable. Weights were then computed for each responding case as the inverses of these estimated response propensities, and weighted estimates of the means on X_2 and X_3 were computed using these weights. Taylor series linearization was used to compute estimates of the standard errors of these estimated means, and corresponding 95% confidence intervals for the means. This approach emulates the current unit nonresponse adjustment for NSFG data.

4) Complete-case (CC) analysis, where analysis is based only on cases with no missing values, with no adjustment of any form for nonresponse, and standard methods for simple random samples are used to compute estimates of means, standard errors, and 95% confidence intervals.

For each simulation, we computed the relative empirical bias (%), root mean squared error (RMSE), 95% confidence / credible interval (CI) coverage, and mean 95% CI width for the estimators of the two means defined by the four approaches above, based on 1,000 samples simulated under alternative missing data mechanisms.

Simulations Based on the Pattern-Mixture Model

The next set of simulations applied the same four analytic approaches to samples simulated from a known PMM. In this case, we expect the PMM approach to out-perform the other three approaches. We aim to contrast the performance of the various approaches depending on the model used to generate the incomplete data.

In the case of low measurement error, we select samples from the following PMM:

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 1.1 \\ 1 \\ 9.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 & 0.25 \\ 0.9 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} \right) \text{ for } m_i = 0;$$

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 2 \\ 2 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 & 0.25 \\ 0.9 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} \right) \text{ for } m_i = 1;$$

$$m_i \sim Bernoulli(\pi_1) .$$

Under this model, nonrespondents have higher values on the two variables of interest (X_2 and X_3), and missingness is a function of values on X_2, which defines the seven parameters restrictions provided earlier. Note the following results under this model:

$$\beta_{32.2} = \sigma_{23}^{(r)} / \sigma_{22}^{(r)} = 0.5 \text{ for } r = 0,1$$

$$\beta_{30.2} = \mu_3^{(r)} - \beta_{32.2}\mu_2^{(r)} = 9 \text{ for } r = 0,1$$

$$\beta_{12.2} = \sigma_{12}^{(r)} / \sigma_{22}^{(r)} = 0.9 \text{ for } r = 0,1$$

$$\beta_{10.2} = \mu_1^{(r)} - \beta_{12.2}\mu_2^{(r)} = 0.2 \text{ for } r = 0,1$$

$$\sigma_{11.2} = \sigma_{11}^{(r)} - \beta_{12.2}^2\sigma_{22}^{(r)} = 1 - 0.81 = 0.19 \text{ for } r = 0,1$$

$$\sigma_{33.2} = \sigma_{33}^{(r)} - \beta_{32.2}^2\sigma_{22}^{(r)} = 1 - 0.25 = 0.75 \text{ for } r = 0,1$$

$$\rho_{13.2} = \frac{\rho_{13}^{(r)} - \rho_{12}^{(r)}\rho_{23}^{(r)}}{\sqrt{1-\rho_{12}^{2(r)}}\sqrt{1-\rho_{23}^{2(r)}}} = \frac{0.25 - 0.9(0.5)}{\sqrt{1-0.81}\sqrt{1-0.25}} \approx -0.53 \text{ for } r = 0,1$$

The measurement error is assumed to be constant across the two patterns (e.g., interviewers in NSFG might make the same types of errors for all sample units). Initial simulation studies have suggested that differential measurement error in auxiliary variables for respondents and nonrespondents may increase the bias in nonresponse-adjusted parameter estimates (Biemer et al., 2011), so the effects of increased measurement error in this simulation study should be considered smaller than effects that would be observed in practice in the case of differential measurement error. The parameter π_1 determines the proportion of cases with missing data on X_2 and X_3, and takes on values of 0.75 and 0.25 (corresponding to high or low unit nonresponse). Notably, the

PMM differs from the selection model in that it does not require assuming a particular form for the missing data mechanism.

The known marginal means of X_2 and X_3 defined by this model are:

$$\mu_2 = 3\pi_1 + (1-\pi_1) = 2\pi_1 + 1;$$

$$\mu_3 = 15\pi_1 + 10(1-\pi_1) = 5\pi_1 + 10.$$

We selected 1,000 samples of size n = 1,000 from this model for each possible value of π_1, and assessed the performance of the same four analysis approaches described earlier. We then repeated the two simulations with the covariance between X_1 and X_2 equal to 0.6 for both patterns (high measurement error), the same values for the other covariance parameters from the first two simulations, the three means for the $m_i = 0$ pattern (respondents) equal to (1.4, 1, 10.5), and the three means for the $m_i = 1$ pattern equal to (2, 2, 11) (to maintain the parameter restrictions).

Results of Simulation Studies

Table 5.1 presents simulation results for each of the four estimation methods (PMM, MI, GW, and CC) under the normal selection model, using alternative values for the parameter defining the non-ignorable missing data mechanism (β), considering two different measurement error scenarios for the fully observed auxiliary variable X_1 (ρ = 0.9 and ρ = 0.6), and setting α = 0.

Table 5.1: Selected simulation results under the normal selection model, with α = 0 in the response propensity model.

ρ	β	Mean RR	Method	$\hat{\mu}_2$ Rel. Bias	$\hat{\mu}_2$ RMSE	$\hat{\mu}_2$ 95% CI	$\hat{\mu}_2$ 95% CI Mean	$\hat{\mu}_3$ Rel. Bias	$\hat{\mu}_3$ RMSE	$\hat{\mu}_3$ 95% CI	$\hat{\mu}_3$ 95% CI Mean

				(%)		Cover.	Width	(%)		Cover.	Width
0.9	2	0.78	PMM	0.02	0.032	0.957	0.130	-0.01	0.038	0.938	0.143
			MI	7.11	0.077	0.379	0.123	1.01	0.108	0.247	0.151
			GW	7.52	0.086	0.574	0.183	1.09	0.116	0.192	0.151
			CC	29.11	0.293	0.000	0.122	1.44	0.148	0.014	0.136
0.9	1	0.70	PMM	-0.18	0.034	0.956	0.132	-0.01	0.038	0.951	0.149
			MI	5.41	0.063	0.612	0.128	0.80	0.088	0.507	0.160
			GW	5.40	0.064	0.774	0.165	0.79	0.087	0.462	0.153
			CC	25.38	0.256	0.000	0.138	1.27	0.132	0.061	0.146
0.9	0	0.50	PMM	-0.24	0.034	0.953	0.138	-0.01	0.045	0.929	0.167
			MI	-0.24	0.034	0.951	0.139	-0.01	0.047	0.944	0.195
			GW	-0.23	0.033	0.994	0.175	-0.01	0.045	0.944	0.176
			CC	-0.22	0.043	0.958	0.176	-0.01	0.045	0.937	0.176
0.6	2	0.78	PMM	-0.25	0.043	0.951	0.167	-0.01	0.039	0.942	0.153
			MI	20.43	0.207	0.000	0.124	1.13	0.118	0.147	0.145
			GW	20.64	0.209	0.000	0.127	1.14	0.119	0.104	0.140
			CC	29.02	0.292	0.000	0.122	1.46	0.150	0.020	0.137
0.6	1	0.70	PMM	0.04	0.045	0.943	0.176	-0.01	0.043	0.934	0.160
			MI	17.37	0.177	0.003	0.139	0.94	0.102	0.342	0.158
			GW	17.38	0.177	0.002	0.143	0.94	0.102	0.302	0.148
			CC	25.68	0.259	0.000	0.138	1.28	0.134	0.089	0.146
0.6	0	0.50	PMM	-0.21	0.053	0.945	0.209	-0.02	0.048	0.948	0.185
			MI	-0.11	0.042	0.943	0.175	-0.01	0.044	0.960	0.196
			GW	-0.15	0.040	0.969	0.176	-0.01	0.043	0.955	0.175
			CC	-0.14	0.045	0.954	0.176	-0.01	0.043	0.955	0.176

NOTES: ρ = corr(X_1, X_2), and defines amount of measurement error in X_1; $\alpha = 0$; β determines dependence of missingness on X_2; PMM = pattern-mixture model estimates based on Bayesian inference approach; MI = multiple imputation estimates after regression prediction and application of Rubin's combining rules; GW = global weighting estimates; CC = complete case estimates; CI = confidence / credible (for PMM) interval

Table 5.2 presents simulation results for each of the four estimation methods (PMM, MI, GW, and CC) under the pattern-mixture model defined earlier. Results are presented for all four combinations of the two measurement error scenarios for the fully observed auxiliary variable X_1 ($\rho = 0.9$ and $\rho = 0.6$) and the two possible values of the parameter defining the proportion of the population that is non-respondents (with values arising from the specified pattern for $m_i = 1$ described earlier).

Table 5.2: Selected simulation results under the pattern-mixture model.

ρ	π_1	Method	$\hat{\mu}_2$ Rel.	$\hat{\mu}_2$ RMSE	$\hat{\mu}_2$ 95%	$\hat{\mu}_2$ 95% CI	$\hat{\mu}_3$ Rel.	$\hat{\mu}_3$ RMSE	$\hat{\mu}_3$ 95%	$\hat{\mu}_3$ 95% CI

			Bias (%)		CI Cover.	Mean Width	Bias (%)		CI Cover.	Mean Width
0.9	0.75	PMM	0.03	0.051	0.948	0.194	-0.02	0.081	0.915	0.287
		MI	-16.35	0.290	0.001	0.227	-4.70	0.471	0.016	0.428
		GW	-8.37	0.170	0.625	0.406	-2.12	0.225	0.281	0.322
		CC	-42.82	0.752	0.000	0.249	-3.82	0.382	0.000	0.249
0.9	0.25	PMM	-0.01	0.036	0.951	0.140	-0.01	0.040	0.929	0.146
		MI	-3.85	0.059	0.714	0.137	-0.72	0.080	0.556	0.153
		GW	-3.84	0.061	0.846	0.176	-0.72	0.079	0.547	0.151
		CC	-19.99	0.253	0.000	0.143	-1.30	0.131	0.086	0.143
0.6	0.75	PMM	-0.02	0.104	0.946	0.415	-0.01	0.090	0.943	0.342
		MI	-29.72	0.524	0.000	0.288	-2.81	0.314	0.081	0.343
		GW	-27.46	0.485	0.001	0.293	-2.41	0.272	0.051	0.279
		CC	-42.77	0.751	0.000	0.248	-3.44	0.380	0.000	0.248
0.6	0.25	PMM	0.16	0.043	0.959	0.174	0.01	0.039	0.959	0.155
		MI	-12.74	0.163	0.010	0.143	-0.82	0.095	0.377	0.152
		GW	-12.74	0.163	0.012	0.150	-0.82	0.094	0.332	0.146
		CC	-19.97	0.252	0.000	0.143	-1.18	0.130	0.072	0.143

NOTES: ρ = corr(X_1, X_2), and defines amount of measurement error in X_1; π_1 defines the proportion of population units with values arising from the model for pattern $m_i = 1$ (non-respondents); PMM = pattern-mixture model estimates based on Bayesian inference approach; MI = multiple imputation estimates after regression prediction and application of Rubin's combining rules; GW = global weighting estimates; CC = complete case estimates; CI = confidence / credible (for PMM) interval

Bias and MSE. When the data were simulated according to a PMM, the Bayesian PMM estimators have the smallest bias and RMSE when missingness depends on the true value, X_2, as expected (Table 5.2). The PMM estimators also perform well in terms of bias and RMSE when the data are simulated from a selection model (Table 5.1). Under the normal selection model and an MCAR mechanism (Table 5.1), the PMM estimators have slightly higher empirical RMSEs under high measurement error, reflecting some loss of efficiency from estimating the nonignorable model parameters. Under both missing data models, the GW and MI estimators have less empirical bias than the CC estimators when the missing data mechanism is non-ignorable, but are still biased, with a bias that increases as dependence of missingness on X_2 and measurement error in X_1 increases. None of the estimators for the mean of the X_3 variable are badly biased in this setting, reflecting the fact that missingness depends on X_2. However, higher proportions of

nonrespondents in the case of the PMM tend to increase the bias and RMSE of the estimators for the mean of the X_3, unlike in the case of the normal selection model. The PMM estimators appear robust to the model generating the missing data and the amount of measurement error in the auxiliary variable.

Confidence / Credible Interval Coverage and Width. Under both missing data models, the coverage of 95% confidence intervals based on the MI, GW, and CC estimators is far below nominal when missingness depends on X_2, and decreases with increased dependence of missingness on X_2 and more measurement error in the auxiliary variable. In contrast, 95% credible intervals based on the Bayesian PMM estimators have close to nominal frequentist coverage in nearly all cases. Interestingly, for higher levels of measurement error (under both missing data models), the mean width of the Bayesian credible intervals based on the PMM estimator tends to be higher than that for the other three estimators, reflecting the fact that increased measurement error in the auxiliary variable increases the uncertainty in the predictive distribution of the missing values.

Similar patterns of results were found for the case where $\alpha = -1$ in the normal selection model (introducing lower response rates). In the cases of non-ignorable missing data mechanisms, the lower response rates simply served to increase the bias and RMSE of the MI, GW, and CC estimators while reducing their coverage. The PMM estimators still performed quite well in the presence of lower response rates, but were once again found to have higher mean confidence interval width in the case of higher measurement error. Despite this, the reductions in bias clearly favored the PMM estimators in this setting,

especially in the case of non-ignorable missing data mechanisms. Interested readers can view Table D.1 in Appendix D for these results.

Applications to Real Survey Data

The Labor Market and Social Security (PASS)[8] *Survey*

We analyzed data from the Labor Market and Social Security (PASS) survey in Germany. The PASS survey is a panel study conducted by the Institute for Employment Research (IAB) in Nuremberg, and collects labor market, household income, and unemployment benefit receipt data from a nationally representative sample of the German population (covering more than 12,000 households annually). This study has a stated purpose of providing "a new database which will allow social processes and the non-intended side-effects of labor market reforms to be assessed empirically."[9] To assist with both stratified sampling and estimation, the PASS survey purchases both continuous and categorical auxiliary variables describing area-level features for sampled households from the German consumer marketing organization Microm[10]. These variables are then linked to the sampled households at the address level, and linking rates are consistently higher than 95%. See Trappmann et al. (2011) for additional details.

We identified continuous variables from the Microm database (available for nearly all sample units) and the PASS survey (Wave 1 respondents in 2006) for analysis. Specifically, 48,250 sampled households had information available on a continuous

[8] PASS: Panel Arbeitsmarkt und Soziale Sicherung; translation = Labor Market and Social Security. The web site for this survey is http://www.iab.de/en/befragungen.aspx.
[9] http://www.iab.de/en/befragungen.aspx
[10] http://www.microm-online.de/Deutsch/Microm/index.jsp

auxiliary variable from Microm measuring the average purchasing power (in Euros) of households in the same city block. This variable followed an approximately normal distribution, and was considered as an error-prone auxiliary proxy (X_1) of reported monthly household income. Monthly household income and area (in square meters) of the housing unit were both measured for 11,969 respondents to the PASS survey in Wave 1 (a 24.8% unweighted response rate)[11].

Monthly household income (log-transformed) was considered as the X_2 variable, and unit nonresponse (on X_2 and X_3) was assumed to be a linear function of this variable. This assumption was supported by a strongly significant association of average household purchasing power with a response indicator, where households with more purchasing power had lower odds of responding. Area of the housing unit (also log-transformed) was considered as the X_3 variable. The correlation between the Microm measure of average purchasing power and the reported household income (log-transformed) was 0.223, suggesting substantial error in the auxiliary proxy (the lowest correlation considered in the simulation studies above was 0.6). The correlation of average purchasing power with log-transformed housing unit area was 0.137, while the correlation of housing unit area and household income was 0.642. Estimates of population means for household income and housing unit area computed using the four methods under study in this paper were exponentiated to return them to their original scales (see Table 5.3).

[11] Unfortunately the base weights and the stratum and cluster codes describing the complex design features of the PASS Wave 1 sample were not available at the time of this analysis. These analyses will be repeated incorporating these complex design features (and using the multiple imputation approach based on the PMM, or PMM-MI, described earlier) at a later date. We illustrate the PMM-MI approach with the analysis in the next section.

Table 5.3: Estimates of mean reported household income and mean housing unit area (in square meters), based on four different nonresponse adjustment methods*.

Variable	Method	Estimated Mean	95% CI	CI Width
Reported Monthly HH Income in Euros (X_2)	CC	1,289.88	1,273.65, 1,306.32	32.67
	GW	1,304.41	1,287.78, 1,321.26	33.48
	MI	1,302.28	1,284.40, 1,320.41	36.01
	PMM	1,597.91	1,522.27, 1,687.58	165.31
Housing Unit Area, Meters Squared (X_3)	CC	73.99	73.38, 74.61	1.23
	GW	74.32	73.70, 74.95	1.25
	MI	74.20	73.63, 74.78	1.15
	PMM	80.99	79.24, 82.91	3.67

* Full sample size: n = 48,250. Respondents: 11,969 (unweighted response rate = 0.248).

Estimates and inferences based on the PMM approach are substantially different, which was evident in the simulations when missingness depended more heavily on the X_2 variable. If the PMM missing data mechanism is assumed to be true, there is evidence of a negative bias in the CC, GW, and MI estimators, which suggests that poorer households are more likely to respond, and that neither the global weighting adjustment nor the item-specific sequential regression imputation is completely correcting for this bias. The widths of the 95% confidence intervals are also larger when using the PMM estimators and the Bayesian approach, which is perfectly consistent with simulation results in the case of higher measurement error in the auxiliary variable and lower response rates (see Table D.1 in Appendix D). Inferences would be more inefficient in this case when using the PMM estimators, but in light of the simulation results above, the large shift in the estimates may be evidence of reduced bias. Finally, we also note that inferences would not vary substantially when using the three alternatives to the PMM estimator, suggesting that the error in the auxiliary variable may be hindering the effectiveness of these more popular nonresponse adjustments.

The Continuous National Survey of Family Growth (NSFG)

We also analyzed data collected from a sample of teenage (age 15-19) persons (n = 3,108) during the first 10 quarters of the recently completed continuous NSFG. The NSFG collects data on fertility, sexual practices, and family characteristics from a continuously released, nationally representative sample of teens and adults ages 15-44 in the United States. A screening interview determines age-eligibility (Lepkowski et al., 2010). The NSFG also collects a variety of paradata or "data collection process information" (Beaumont, 2005; Couper, 1998; Couper and Lyberg, 2005) both *prior* to screening interviews (e.g., number of attempted calls, previous statements made by potential respondents, many-unit housing structure, neighborhood characteristics) and *during* the screening interviews (e.g., age, presence of a member of the opposite sex, single-person household). The interviewers working in the NSFG (all of whom are female) judge the perceived current sexual activity of a selected respondent *immediately upon completion* of the screening interview, so the aforementioned paradata would be available to the interviewer at the time of making a judgment. In total, there were 772 sampled teenagers (24.8% of the full sample) providing responses on an NSFG variable (X_3) measuring age at first sexual activity. All 3,108 teenagers had sexual activity judgments available and provided an indicator of current sexual activity in the survey.

Auxiliary variables are often categorical in survey research, especially in the case of interviewer observations. Given that the present study describes methods for trivariate *normal* incomplete data, we considered replacing the error-prone interviewer judgment of current sexual activity (available for all 3,108 sample units) with a continuous proxy

variable (e.g., Andridge and Little, 2009). This proxy variable was the estimated linear predictor from a logistic regression model for the binary interviewer judgment, with the aforementioned paradata as predictor variables. This computed proxy variable was the auxiliary variable X_1, available for the full sample. The "true" values for this continuous proxy (the X_2 variable), which were also available for the full sample, were the estimated linear predictors from a logistic regression model for the binary *respondent report* of current sexual activity, using the same paradata as predictors. Values on X_2 were deleted for the non-respondents to X_3 to simulate the unit nonresponse case being studied here.

The estimated correlation of X_1 and X_2 was strong in this sample ($r = 0.91$, $p < 0.001$, $n =$ 3,108), indicating low measurement error in the proxy judgment. The correlation of X_1 and X_3 was much weaker and negative, as expected ($r = -0.09$, $p = 0.01$, $n = 772$). A significant positive relationship of known values on X_2 with a response indicator for X_3 (and thus X_2) from a fitted logistic regression model provides support for our assumption that missingness depends on values of the true auxiliary variable (X_2). Those with higher probabilities of being sexually active also had higher probabilities of responding on both survey items.

The NSFG employed a complex multistage sampling design (see Lepkowski et al., 2010). All analyses therefore incorporated available base sampling weights for unbiased estimation of the means, and computed linearized standard errors using the sampling error stratum and cluster codes available in the public-use NSFG data set. In the global weighting (GW) method, available base sampling weights for the 772 responding sample

units were multiplied by the inverse of the estimated response propensity based on an estimated logistic regression model predicting a response indicator for the 3,108 sample units with the computed proxy variable X_1. In the MI approach, the complex sampling features were accounted for in each complete-case analysis and $B = 20$ imputed data sets were analyzed. The PMM approach utilized the multiple imputation approach to inference described earlier (denoted by PMM-MI), again incorporating the complex sampling features in each complete-case analysis. An inverse logit transformation was used post-estimation to transform estimates and confidence limits for the continuous proxy of sexual activity back to a probability scale. Table 5.4 presents estimates of (and 95% confidence intervals for) the population means for the X_2 and X_3 variables, based on three of the four alternative methods studied in the simulations (CC, GW, and MI) and the PMM-MI approach.

Table 5.4: Estimates of selected means in the NSFG teenage population, based on four different nonresponse* adjustment methods.

Variable	Method	Estimated Mean	95% CI	CI Width
Currently Sexual Active (X_2)	CC	0.572	0.565, 0.579	0.014
	GW	0.548	0.541, 0.556	0.015
	MI	0.553	0.549, 0.557	0.008
	PMM-MI	0.550	0.546, 0.553	0.008
Age at First Intercourse (X_3)	CC	15.767	15.521, 16.012	0.491
	GW	15.871	15.630, 16.111	0.481
	MI	15.856	15.699, 16.013	0.314
	PMM-MI	15.859	15.733, 15.986	0.253

* Full sample size: n = 3,108. Respondents: 772 (unweighted response rate = 0.248).

Inferences for the proportion of teenagers sexually active would be different when using the GW, MI, or PMM-MI approaches instead of the CC approach, each of which indicate that the proportion of teenagers sexually active is lower than suggested by the CC

analysis. The 95% CI for the proportion based on the PMM-MI approach includes the point estimates suggested by the GW and MI approaches, and inferences based on these three approaches are presumably similar given the high correlation of X_1 and X_2 in this case ($r = 0.91$). Inferences for the mean age at first sexual activity would be similar when using the four alternative approaches, and this is also not entirely surprising given the low correlation of X_3 with both X_1 and X_2.

As is evident from their forms, the PMM and PMM-MI estimators will reduce bias when 1) missingness is substantially related to the underlying true value (not the case in this example, given that missingness on X_2 depended on whether a respondent provided an answer to X_3, and these two variables had a low correlation); 2) the auxiliary proxy has substantial measurement error, making the MAR adjustment inadequate (also not the case in this example, where X_1 and X_2 had a correlation of 0.91); and 3) the missing data rate is high. If the measurement error in the auxiliary proxy is so large that the correlation between the proxy and the true variable is low (as was the case in the PASS example), then bias reduction will come at the expense of increased variance (also noted in the PASS example).

Generalization to Include Other Fully Observed Auxiliary Variables

In practice, there will often be other auxiliary variables that we wish to include as predictors in models for imputing missing values. Suppose that in addition to the data in Figure 5.1 there is a set of k auxiliary variables C that are fully recorded and not subject to measurement error (including a vector of 1s for the intercept term), and that

missingness of X_2 and X_3 is assumed to depend on both X_2 and C. Then, for the missing data pattern $m_i = r$, we assume the following generalization of the model described earlier. Conditional on values c_i of the auxiliary variables C,

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \beta_{1c.c}^{(r)} c_i \\ \beta_{2c.c}^{(r)} c_i \\ \beta_{3c.c}^{(r)} c_i \end{pmatrix}, \begin{pmatrix} \sigma_{11.c}^{(r)} & \sigma_{12.c}^{(r)} & \sigma_{13.c}^{(r)} \\ . & \sigma_{22.c}^{(r)} & \sigma_{23.c}^{(r)} \\ . & . & \sigma_{33.c}^{(r)} \end{pmatrix} \right) \equiv N_3 \left(\beta_{c.c}^{(r)} c_i, \Sigma_c^{(r)} \right), \tag{5.10}$$

a trivariate normal distribution with $3(k + 1) + 6$ parameters. In addition, the marginal distribution of m_i given z_i under this PMM is $m_i \mid c_i \sim Bernoulli\left(\pi_1(c_i), \gamma\right)$.

The following parameters are identified from the observed data in Figure 5.1:

$$\gamma, \beta_{1c.c}^{(0)}, \beta_{2c.c}^{(0)}, \beta_{3c.c}^{(0)}, \sigma_{12.c}^{(0)}, \sigma_{13.c}^{(0)}, \sigma_{23.c}^{(0)}, \sigma_{11.c}^{(0)}, \sigma_{22.c}^{(0)}, \sigma_{33.c}^{(0)}, \beta_{1c.c}^{(1)}, \text{ and } \sigma_{11.c}^{(1)}.$$

The following $2k + 7$ parameters are not identified:

$$\beta_{2c.c}^{(1)}, \beta_{3c.c}^{(1)}, \sigma_{12.c}^{(1)}, \sigma_{13.c}^{(1)}, \sigma_{23.c}^{(1)}, \sigma_{22.c}^{(1)}, \text{ and } \sigma_{33.c}^{(1)}.$$

Let $\beta_{ic.c}^{(r)}$ denote the coefficients for the set of auxiliary variables C in the linear regression of variable i on C for pattern r. Also, let $\sigma_{ij.c}^{(r)}$ denote the residual covariance (or variance, if $i = j$) of variables i and j, given C, for pattern r. The assumption that missingness of X_2 and X_3 depends on X_2 and C implies that the distribution of X_1 and X_3 given X_2 and C is the same for complete and incomplete cases, yielding $2k + 7$ parameter restrictions. Hence the model is just-identified (as described earlier).

The PMM can then be estimated as follows. The identified parameters of the conditional distribution of (X_1, X_2, X_3) given C in each pattern are computed as before, with the regression coefficients on C computed by applying OLS regression to the two patterns.

The other parameters are similar functions of the identified parameters given earlier, except that the expressions condition on the auxiliary variables *C*. The ML estimates are computed as follows, given the notation below (where *C* includes the column of 1s used for the intercept terms in the models):

$\hat{\beta}_{1c.c} = \text{OLS regression coefficients of } X_1 \text{ on } C, \text{ all data}$
$\sigma_{11.c} = \text{ Residual variance of } X_1 \text{ given } C, \text{ all data}$
$b_{jc.c} = \text{OLS regression coefficient of } X_j \text{ on } C, \text{ complete cases}, j = 1,2,3$
$b_{j2.2c} = \text{Coefficient of } X_2 \text{ from OLS regression of } X_j \text{ on } C \text{ and } X_2, \text{ complete cases}, j = 1,3$
$s_{jk.c} = \text{ Covariance of } X_j, X_k \text{ given } C, \text{ complete cases}$

$$\hat{\beta}_{2c.c} = b_{2c.c} + \frac{\hat{\beta}_{1c.c} - b_{1c.c}}{b_{12.2c}}$$

$$\hat{\beta}_{3c.c} = b_{3c.c} + b_{32.2c}\frac{\hat{\beta}_{1c.c} - b_{1c.c}}{b_{12.2c}}$$

$$\hat{\sigma}_{12.c} = s_{12.c} + \frac{\hat{\sigma}_{11.c} - s_{11.c}}{b_{12.2c}}$$

$$\hat{\sigma}_{22.c} = s_{22.c} + \frac{\hat{\sigma}_{11.c} - s_{11.c}}{b_{12.2c}^2}$$

$$\hat{\sigma}_{13.c} = s_{13.c} + b_{32.2c}\frac{\hat{\sigma}_{11.c} - s_{11.c}}{b_{12.2c}}$$

$$\hat{\sigma}_{23.c} = s_{23.c} + b_{32.2c}\frac{\hat{\sigma}_{11.c} - s_{11.c}}{b_{12.2c}^2}$$

$$\hat{\sigma}_{33.c} = s_{33.c} + b_{32.2c}^2\frac{\hat{\sigma}_{11.c} - s_{11.c}}{b_{12.2c}^2}$$

A sequence of draws from the posterior distribution of the identified parameters in this case can be computed assuming noninformative priors for the identified parameters, in the same way as described earlier, and the draws of the identified parameters can be plugged in to the estimates above to simulate draws from the posterior distribution of

these parameters. Bayesian inference as discussed earlier for the trivariate case can then proceed in a straightforward manner.

Discussion

This study has demonstrated the desirable properties of PMM estimators for population means in a survey context, where a fully observed continuous auxiliary variable has been measured with error on all sample units, true values on the auxiliary variable (along with other continuous survey variables of interest) are measured on survey respondents, and missingness arises as a function of true values on the auxiliary variable. Simulation studies suggest that PMM estimators computed using a Bayesian approach have reduced bias, reduced RMSE, 95% credible sets with nominal frequentist coverage, and reasonable mean credible set width relative to weighted, multiple imputation, and complete-case estimators in the presence of a non-ignorable missing data mechanism. The PMM estimators are also shown to be robust to varying levels of measurement error in the fully observed auxiliary variable, with the exception of increased credible set width in the case of higher measurement error.

In general, increased measurement error in the auxiliary proxy lead to increased empirical bias, increased RMSE, and poorer confidence interval coverage of the competing (and more widely used) estimators. Given that these properties of the more popular estimators may already be poor in the presence of a non-ignorable missing data mechanism, measurement error can exacerbate this problem, and pattern-mixture model estimators are shown to be robust in this regard. We also found the PMM estimators to be robust to the

model generating the missing data, as these estimators performed equally well when missing data were generated under a normal selection model.

Analyses of real data from two large area probability sample surveys in the United States and Germany illustrate the use of the item-specific PMM estimators in practice. Analyses of data from the NSFG in the United States illustrate how the PMM approach can be implemented when analyzing survey data collected from a complex sample. In general, the forms of the PMM estimators indicate situations where one can expect the most bias reduction: 1) missingness is substantially related to the underlying true value; 2) the auxiliary proxy has substantial measurement error, making the MAR adjustment inadequate; and 3) the missing data rate is high. As shown in the simulation studies, if the measurement error in the auxiliary proxy is large enough that the correlation between the proxy and the true variable is low, then bias reduction will come at the expense of increased variance. There was evidence of this in the analyses of the PASS survey data from Germany, but more general conclusions about the estimators should not be drawn from these two very specific examples, where the true underlying missing data mechanisms were not known.

Results from the analyses of real survey data were found to be consistent with empirical results from the simulation studies, but additional applications are certainly needed in various survey contexts to further study the effectiveness of the proposed estimators. The non-ignorable missing data mechanisms studied in the simulations may not be entirely realistic in practice, where the use of several auxiliary variables associated with key

survey variables in nonresponse adjustments may result in missing data mechanisms that could be considered ignorable. These variables tend to be rare in survey practice, however, which increases the attractiveness of the PMM estimators studied here.

The possibility arises that estimates of the mean (and confidence intervals for the mean) of the true auxiliary variable measured in the survey may vary depending on the third survey variable being analyzed when using the trivariate normal PMM approach proposed in this paper. This possibility was investigated using the NSFG data, where a second survey variable containing self-reports of general health was analyzed using the PMM methodology. Estimates of the proportions for sexual activity and their confidence intervals did not vary substantially from the estimates found when using age at first intercourse as the survey variable, but this possibility needs further investigation.

There are many possible extensions of this work. This work only considered a single auxiliary variable measured with error, and extensions to more than two such variables (as outlined above) would be useful. The second real data example in this study considered methods for replacing a binary auxiliary variable with estimated linear predictors based on a vector of auxiliary variables, but general extensions of the PMM estimators to accommodate a vector of fully observed auxiliary variables of different types (binary, normal, etc.) would also be important (e.g., Little and Wang, 1996). Recent work has also developed methods for selecting the best vectors of fully observed auxiliary variables to be used in calibration estimators for reducing nonresponse bias (Sarndal and Lundstrom, 2010), and given the selection of the "best" vector for a given

survey variable, the performance of PMM estimators based on this vector of auxiliary variables could be compared to that of the calibration estimators. Further extensions might include development of PMM estimators for additional binary variables measured in the survey, given the importance of binary outcomes in survey research, and work is currently ongoing in this area (Andridge and Little, 2009). Finally, we also assumed that there was no measurement error in the survey variables measured for respondents, and the impact of error in these variables on the methods discussed in this study also deserves future research attention.

The results in this chapter suggest that PMMs may be a valid alternative to the global weighting and multiple imputation techniques that are widely used in multipurpose sample surveys, especially when the auxiliary variables used to construct nonresponse adjustments are measured with error and non-ignorable missing data mechanisms are suspected with respect to true values on those auxiliary variables collected in the survey. Although separate models would need to be formulated for each survey variable in this framework, development of software to facilitate this task should enable more survey analysts to use these methods and produce estimates with potentially higher quality. R functions enabling applications of the PMM estimators proposed in this study to real survey data are provided in Appendix C.

Chapter VI

Conclusions and Discussion

The results of the studies presented in Chapters 3 through 5 have important practical implications for the data collection and estimation strategies that will be used in future surveys. This concluding chapter considers these implications, along with directions for future research in this area.

Synthesis of Study Results

For a given survey context (e.g., health, politics, fertility, drug use, social views, etc.), survey managers need to carefully consider correlates of both key survey variables and survey cooperation that interviewers will be able to observe and record. Ideally, these auxiliary variables observed by the interviewers will also be measured in the survey interview, providing the validation data needed to monitor the accuracy of the observations in future empirical studies. Consultation with subject matter experts having theoretical knowledge about *observable* predictors of key survey variables and predictors of response propensity will be essential for this task, and will allow survey methodologists, sociologists, and psychologists to work together and further interdisciplinary research aimed at improving the quality of survey estimates.

Survey managers can use the error properties reported for the two NSFG interviewer observations in Chapter 3 as benchmarks for future data collections that will require interviewers to make similar respondent- and household-level judgments. Importantly, these observations were found to be between 70 and 80% accurate (Chapter 3), and substantial interviewer variance was found in the accuracy of both observations (Chapter 4). Unfortunately, the accuracy of the respondent-level judgment could only be computed based on respondent reports, and the accuracy of judgments for non-respondents could not be computed. Future methodological studies of the errors in these types of interviewer judgments should consider alternative sources of validation data, including external data sources that might be linked to data sets containing the interviewer observations (e.g., administrative records). Assuming no major issues with record linkage, these alternative data sources would enable study of the errors in the interviewer judgments for both respondents and nonrespondents, which was not possible in the study presented in Chapter 3.

Chapter 3 also showed that despite significant associations of the NSFG interviewer observations with both response indicators and key NSFG variables, nonresponse adjustments incorporating the two observations did not tend to shift descriptive estimates substantially. This may have been due to high response rates for the main NSFG interview conditional on a completed screener (~81%). Another explanation may have been the complex associations of the interviewer observations with all of the other auxiliary variables (including a variety of paradata) used in the response propensity models to compute the nonresponse adjustments. Initial work by Kreuter and Olson

(2011) has demonstrated that the effectiveness of nonresponse adjustments based on two auxiliary variables depends heavily on the associations of the two auxiliary variables with each other, in addition to their associations with response propensity and the key survey variables. Future research in this area needs to develop more general results and expectations based on these complex associations when more than two auxiliary variables are used to compute the nonresponse adjustments.

The errors in the NSFG interviewer observations may have also affected the quality of the nonresponse adjustments. A small simulation study presented in Chapter 3 (using real NSFG data to define a hypothetical population) showed that weighting class adjustments using predicted response propensities based on the "true" value of an auxiliary variable eliminated the bias introduced by a missing at random (MAR) mechanism, where missingness on key survey variables was a function of the true values of the auxiliary variable. Under the same mechanism, weighting class adjustments using the interviewer observations on the auxiliary variable led to estimates with increased MSE compared to both complete case estimates and adjusted estimates based on the true auxiliary variable. This was the first empirical work to date that has reported the negative implications of errors in interviewer observations for the effectiveness of weighting class adjustments for nonresponse. Importantly, Biemer et al. (2011) also found that higher levels of disposition-dependent error in interviewer-reported call records (i.e., different error rates in the call records for completed interviews, refusals, non-contacts, etc.) can substantially bias estimates of population proportions based on *callback models* (Biemer et al., 2010) when the different groups defining the proportion have different probabilities of

responding. Collectively, the results of these two studies provide motivation for future research examining improved estimation methods that are robust to the errors in these auxiliary variables, and/or design strategies for minimizing the errors in interviewer observations.

Simulation results presented in Chapter 5 echoed the simulation results in Chapter 3, providing further evidence of the significant negative impact of errors in auxiliary variables on the quality of popular nonresponse adjustments. In general, regardless of the model from which trivariate normal incomplete data were generated, higher levels of error in a fully observed auxiliary variable were found to increase bias and MSE while decreasing confidence interval coverage for three alternative estimators of means for the variables with incomplete data (complete case estimators, estimators based on "global" response propensity weights and multiple imputation estimators). Interestingly, Bayesian inferences based on pattern-mixture model (PMM) estimators were found to be robust to different levels of measurement error in the auxiliary variable, regardless of the model used to generate the incomplete data or the missing data mechanism. When combined with the results of the simulations in Chapter 3, these findings lend empirical support to theoretical results presented by Lessler and Kalsbeek (1992, Chapter 8, p. 190), suggesting that decreased homogeneity in weighting classes has the potential to cause adjusted estimates to have more bias than unadjusted estimates.

Although interviewer observations may be used for a variety of purposes in survey research, they will never be entirely error-free, and future research of their utility for

nonresponse adjustments needs to focus on the development of consistent estimators that can simultaneously handle the potential biases introduced by missing data and errors in the observations. The "built-in" calibration of the error-prone auxiliary variable that defines the PMM estimators in Chapter 5 (where "true" values of auxiliary variables are collected from survey respondents) is likely one of the key properties of this estimator. Other recent work focusing on this problem has considered a Monte Carlo EM methodology for model estimation, where the likelihood for a set of data is partitioned into a model for the response of interest (e.g., a generalized linear mixed model), a model for the measurement error in the covariates (e.g., a linear regression model), and a model for the missing data mechanism (e.g., a logistic regression model) (Yi et al., 2011). Unfortunately, many variable-specific methods like this ultimately require untestable assumptions, and require specific models for each survey item (which would certainly increase analyst burden in a multipurpose survey). Because of these problems, sensitivity analyses considering different assumptions for existing methods may be best for practice (Yi, 2011). Future research focusing on the development of software enabling these sensitivity analyses, in combination with evaluation of the performance of these alternative methods in the case of nonresponse adjustment of survey estimates using error-prone auxiliary variables, will be important for practicing survey researchers.

Ultimately, however, secondary analysts of public-use survey data sets will not have access to information for nonrespondents, and future research needs to also consider improved methods for computing nonresponse adjustments for responding cases. In general, methods for correcting the bias in predicted response propensities that may be

introduced by errors in auxiliary variables (Stefanski and Carroll, 1985) should be employed when response propensity models are used to compute "global" nonresponse adjustments in multipurpose surveys. This may be done in practice by correcting the bias in estimates of parameters in logistic regression models that are used to compute response propensities, or by some method of "smoothing" the predicted response propensities. Alternative methods for correcting the bias in estimates of logistic regression model parameters introduced by errors in the covariates have been proposed in the literature. These include misclassification simulation-extrapolation, or MC-SIMEX, for binary auxiliary variables measured with error in logistic regression models (Kuchenhoff et al., 2006), nonparametric maximum likelihood estimation methods that relax assumptions of normality for true covariates and the measurement errors (Rabe-Hesketh et al., 2003), and regression calibration, or the use of large validation studies relating the error-prone measures on auxiliary variables to available true measures to reduce bias in the estimated coefficients (Bosner et al., 1990; Spiegelman et al., 2000). These methods all require access to some form of validation data for the error-prone auxiliary variables, which may not be feasible in survey practice unless the auxiliary variables are also measured for respondents and these responses are assumed to be "truth" (similar to the case of the PMM estimators in Chapter 5).

Chapter 5 uses both simulation studies and analyses of real survey data to illustrate how PMM estimators have desirable properties compared to estimators based on global survey weights (possibly incorporating unit nonresponse adjustments) and multiple imputation estimators assuming missing at random (MAR) mechanisms for individual survey items,

when missing data mechanisms are non-ignorable with respect to values on a key variable measured in the survey and error-prone measures of that variable are available for respondents and nonrespondents alike. Importantly, Chapter 5 also illustrates a method for imputing missing values based on the PMM and then performing multiple imputation analyses incorporating complex sample design features. This methodology should appeal to analysts of complex sample survey data who do not have access to information for nonrespondents. Survey agencies suspecting non-ignorable missing data mechanisms for key survey variables and working with error-prone auxiliary variables available for the full sample could impute missing values based on PMMs as shown in Chapter 5, and then release multiple imputed data sets with complex sample design information to the public, along with instructions for performing multiple imputation analyses of the data. Appendix C provides R code for performing these kinds of PMM-based multiple imputation analyses.

Survey agencies conducting multipurpose surveys may be reluctant to use the item-specific PMM approaches assuming non-ignorable missing data mechanisms, continuing to compute global survey weights that incorporate nonresponse adjustments based on potentially error-prone auxiliary variables. These agencies should form nonresponse adjustment cells based on recoded values of predicted response propensities (e.g., deciles; see Little, 1988, Section 5, or Beaumont, 2005) and then compute the overall response propensity for all respondents within a particular cell (and use this response propensity for all respondents in that cell when computing the nonresponse adjustments). This may be an effective strategy for eliminating the effects of error in the auxiliary variables on

the predicted response propensities (Stefanski and Carroll, 1985). This is often done in practice to reduce the variance of the weights that arises from large variance in the response propensity adjustments, which can increase the variance of sample estimates (Little, 1988). Other methods of smoothing the weights after performing nonresponse adjustments (e.g., Little, 1986) may also be useful for reducing the impact of error in the auxiliary variables on the adjustments. Using an instrumental variable (IV) approach to replacing error-prone values on an auxiliary variable with predicted values based on relevant instruments (e.g., West, 2010a) could also serve well to reduce the bias introduced by errors in the auxiliary variable. Future research should continue to simultaneously evaluate a) the amount of bias and variance in survey estimates introduced by predicted response propensities computed using auxiliary variables with varying levels of error, and b) the impact of the alternative techniques discussed above on reducing the bias and variance in the estimates.

Survey researchers working with error-prone auxiliary variables need to couple improved estimation strategies with effective design strategies targeted at minimizing the error in the auxiliary variables. Chapter 4 presented novel results from a qualitative analysis examining justifications provided by 45 NSFG interviewers for their judgments of current sexual activity. Nearly 4,000 justifications provided by these interviewers were coded into 13 binary indicators of specific cues mentioned in the justifications (e.g., mention of relationship status in the screening interview), in addition to the number of words typed for each justification in the CAPI instrument. These 14 variables were aggregated into percentages and means for each of the 45 interviewers, and then

standardized. A cluster analysis of the 14 standardized variables revealed four distinct clusters of interviewers in terms of their justification tendencies, suggesting that different interviewers did in fact tend to use different strategies for making their judgments. When included as an interviewer-level covariate in multilevel models for the accuracy of the current sexual activity judgments, the categorical cluster identifier was found to have a marginal impact on the odds of a false negative judgment relative to a correct judgment (holding a variety of other interviewer-level and respondent-level predictors fixed). Specifically, the cluster of interviewers defined largely by justifications based on age and observations of household characteristics was found to have significantly reduced odds of a false negative judgment.

Given the interviewer variance in accuracy of the sexual activity and young children judgments reported in Chapter 4, the finding of variance in observational strategies and judgment accuracy based on those strategies may have important implications for future interviewer training purposes. If standardized observational methods are used when making a judgment, interviewer variance in judgment accuracy may be decreased. Future replications of this qualitative study are needed in other survey contexts to see if this phenomenon of variance in observational techniques persists, and evaluations of training efforts designed to decrease interviewer variance in accuracy based on standardizing observational techniques (according to techniques found to be associated with increased accuracy) are also needed. The need for harmonization of observational techniques across interviewers has been suggested previously by Stoop et al. (2010, p. 301), and the European Social Survey (ESS) is taking steps to provide interviewers with photographs

of various environmental settings and discuss appropriate observations in each case during interviewer training (Stähli, 2010).

In addition, providing systematic feedback to interviewers on the accuracy of their judgments and the techniques being used to make them may also help to improve judgment accuracy (Heuer, 1999, Ch. 5). Provided that interviewers are not making (standardized) observations according to a specific type of training, future research should also consider interviewer-specific trends in justifications used for their judgments. Olson and Peytchev (2007) found that interviewer behaviors (i.e., length of survey administration) and perceptions do in fact change over the course of a data collection, and observational strategies may change over time as well in the absence of standardized training.

The multilevel analyses of factors impacting the accuracy of interviewer observations in the NSFG (Chapter 4) provide the first empirical evidence regarding relationships of interviewer- and respondent-level predictors with the accuracy of interviewer judgments in a personal interview setting, and show that multilevel modeling can be used to identify predictors of observation accuracy. In general, the accuracy of the household-level NSFG interviewer observation (on the presence of young children) was found to be a function of respondent-level features and interactions between respondent- and interviewer-level features, but not a function of interviewer-level features *alone*. In the case of the person-level judgment (current sexual activity), accuracy was found to be a function of respondent- and interviewer-level features, but not interactions between these features.

These analyses identified respondent, household, neighborhood, and area characteristics that made judgments of the presence of young children and current sexual activity more difficult (e.g., living in an urban area), and should motivate future studies of improved observational techniques for more difficult subsets of respondents (e.g., analyses of predictors of sexual activity reports for respondents living in urban areas that could assist interviewers with these more difficult judgments). Linked information from commercial databases may also serve as reasonable substitutes for interviewer observations in areas where observations are more difficult, but this is an area for future research. The findings regarding the accuracy of the household-level observation in Chapter 4 were fully consistent with results from the visual cognition literature, which has shown that observational task difficulty (as opposed to individual differences) has been shown to influence accurate detection of other objects (Simons and Jensen, 2009). In addition, consistent with a previous study by McCulloch et al. (2010), increased levels of interviewer experience (in terms of days working on the current cycle of NSFG) were found to result in more systematic false negative judgments of presence of young children, and more systematic false positive judgments of current sexual activity. Given evidence of interviewers changing their behavior over the course of a data collection (Olson and Peytchev, 2007), these results point to the need for continuous monitoring of the quality of interviewer observations over the course of a data collection.

The analyses in Chapter 4 also revealed cross-level interactions between respondent- and interviewer-level predictors that influenced the accuracy of housing unit observations.

For example, certain respondent-level predictors were found to decrease observation accuracy for interviewers with certain characteristics but not others, and this has important implications for the training of interviewers on these observational techniques. Specifically, survey managers need to have consistent discussions with interviewers having specific characteristics that were found to minimize negative effects of area-level predictors on accuracy, and identify the observational strategies used in these specific situations. These techniques can then be shared with all interviewers in more general interviewer training sessions. Current research on inattentional blindness suggests that when observers adopt an *attentional set*, or ready themselves to receive a specific type of information, they are more likely to notice other objects while performing an otherwise demanding observational task (Most et al., 2005). Given this research, conversations with interviewers who make more accurate observations when faced with certain difficulties (e.g., many-unit buildings) may reveal attentional sets (or cues to look for that indicate the presence of children in these environments) that would be useful to share with all interviewers in the training sessions.

The multilevel models estimated in Chapter 4 also indicated that a large amount of unexplained between-interviewer variance in the accuracy of both NSFG observations (and the effects of respondent-level features on accuracy) remained after accounting for all of the available interviewer-level features. Multilevel models provide a tool for estimating proportions of unexplained variance in outcomes that are due to factors at a certain level of a data hierarchy; for example, the varying observational strategies used by interviewers were found to explain approximately 16% of the variance among

interviewers in the odds of making a false negative judgment relative to a correct judgment on current sexual activity. The results in Chapter 4 clearly suggest that unexplained variance in accuracy rates remain among interviewers, even after accounting for several respondent- and interviewer-level characteristics. Future research therefore needs to focus on further identification of factors resulting in this unexplained variance.

Variance among interviewers in perceptive ability may be one source of the unexplained variance in judgment accuracy found in Chapter 4. Given findings from the social psychology literature (see Chapters 2 and 4), future interviewer training methods for making quick judgments more accurate might also incorporate tests of perception and nonverbal decoding ability. Past research on person perception has used the Interpersonal Perception Task (IPT) to evaluate judgment accuracy (Ambady et al., 1995; Patterson and Stockbridge, 1998; Patterson et al., 2001; Smith et al., 1991), and survey organizations might consider adding this as an evaluative tool for interviewers. Per Ambady et al. (1995, p. 527), "…in situations where fairly rapid judgments have to be made regarding others…, good judges of behavior might be identified by their performance on tests of nonverbal decoding ability." The NSFG does not give interviewers these tests at present, but other surveys requiring interviewers to make these types of observations might consider results from this literature, and test potential interviewers on their perceptive ability. These test scores could then be used as another potential interviewer-level predictor of accuracy in these types of multilevel models. If testing interviewers on perceptive ability is not reasonable from a cost efficiency point of view, or does not result in a fair trade-off between cost of training and data quality, questions asking interviewers

to self-report their perceptive ability on employee questionnaires may serve as reasonable proxies, but this is an area for future research.

Another interesting finding from the social psychology literature involves the role of a judge's mood in the accuracy of their judgments, with judges found to be sad or more depressed having less accurate judgments in thin-slice scenarios (Ambady and Gray, 2002). Many surveys request interviewers to record their opinions of the quality of an interview after the completion of an interview, including any potential problems. In surveys like the NSFG which conduct initial screening interviews, some main interviews may be completed *immediately* after the screening interview. Post-interview questions in these scenarios could therefore include questions about the mood of the interviewer at the onset of the main interview, when the screening interview judgment would typically be recorded. Future studies could then determine whether the reported moods in these "immediate" interview situations have any influence on judgment accuracy.

Chapter 4 also presented results evaluating the effectiveness of a novel methodology that provides interviewers with significant (and observable) predictors of the primary features that they are trying to observe, to see whether this technique improves the accuracy of their observations. In the NSFG, all interviewers were provided with this predictive information for actual respondent reports of current sexual activity (based on the first 14 quarters) in the last two quarters of data collection. In the multilevel analyses of accuracy on the current sexual activity judgment presented in Chapter 4, a binary respondent-level indicator of measurement in these two quarters was found to significantly reduce the odds

of making a false positive judgment (which is a documented problem with these judgments; see Chapter 3) as opposed to a correct judgment. Importantly, this relationship was found when controlling for days of experience working on NSFG (Cycle 7), along with several other respondent- and interviewer-level predictors. These results combine with existing theory from social psychology to provide empirical support for the effectiveness of this type of design strategy for increasing observation accuracy.

Unfortunately, this provision of predictive information to all interviewers in the last two quarters of NSFG Cycle 7 was not a randomized intervention. In an effort to generalize the results presented in Chapter 4, a randomized experiment has been embedded in a new economic survey in Germany to evaluate the effectiveness of this methodology more generally in an original data collection. If additional studies can demonstrate that this technique is shown to be effective at improving the accuracy of interviewer observations, survey managers will be able to study past survey data and identify predictors of the auxiliary variables that interviewers are attempting to observe that are available to the interviewers when they are making their judgments (e.g., paradata on previous calling efforts, neighborhood features, etc.). Interviewer training should then focus on techniques for detecting these available predictors prior to making judgments (Funder, 1995).

Future Directions for Survey Methodology

The research presented in this dissertation has focused on interviewer observations collected in personal interview surveys only. Research into alternative methods of collecting observations on all sampled households in web, mail, or telephone surveys is

certainly needed, given that auxiliary variables available for all sample units tend to be extremely rare when using these modes of data collection. For example, satellite imagery from software products such as Google Earth could be used to discern characteristics of sample households in a mail survey, enabling the collection of observations on both responding and non-responding households for eventual nonresponse adjustments.

Future research also needs to focus on the error properties of auxiliary variables available in large commercial databases (e.g., MSG, Experian and Axciom in the United States, or Microm in Germany). Survey researchers are currently linking the auxiliary variables available in these commercial household databases to sampling frames compiled from listing operations, address-based sampling frames, or other frames developed for non-personal-interview modes, in an effort to enrich the information available on the frames. These auxiliary variables may then be used for many methodological purposes, including stratification of samples, screening operations, and nonresponse adjustment of survey estimates. However, this practice may be dangerous if the auxiliary variables available in these databases have poor quality, and more research is certainly needed in this area.

Initial work in this area has in fact suggested that the data in these commercial databases may be of very poor quality (Daily et al., 2008; Hubbard and Lepkowski, 2009; Pickering et al., 2003, p. 8; Yan et al., 2011), and future work needs to consider the implications of this reduced quality for current survey methodologies. Future research should compare the quality of both interviewer observations on a particular feature and linked auxiliary data on the same feature from a commercial database (e.g., interviewer observations on

the presence of young children under the age of 15, versus indicators of young children under the age of 15 in a household based on available commercial data for a household). This type of research could help survey methodologists determine the cost-error tradeoffs between having interviewers collect auxiliary variables and purchasing auxiliary information from commercial vendors. Ultimately, the implications of errors in both sets of auxiliary variables for nonresponse adjustments, screening operations, sampling efficiency, etc. should be weighed against the costs of collecting both types of variables.

Appendices

Appendix A: Additional Exploratory Analyses from Chapter 4

Additional Exploratory Analyses: Accuracy of Young Children Judgments. Several additional exploratory analyses were conducted to examine potential cross-level interactions of respondent-level predictors with coefficients that were initially found to significantly vary across a *reduced* number of interviewers (due to a lack of consistent within-interviewer variance in the predictors) with interviewer-level features. For one respondent-level predictor at a time [as recommended by Hosmer and Lemeshow (2000) for testing interactions in logistic regression models], all of the interviewer-level predictors were added to the Level 2 equations for the coefficients of that respondent-level predictor in the two logit functions (while retaining the interviewer-level predictors in the Level 2 equations for the intercepts). Cross-level interactions of the urban area indicator with interviewer-level features could not be estimated, for reasons that HLM attributed to multicollinearity; this was likely due to high associations of the urban area indicator (computed from the type of PSU where a respondent resided) with the interviewer-level indicators of types of PSUs worked by the interviewers.

Several significant interactions were found between the respondent-level indicator of living in a residential segment with interviewer-level features. Specifically, a significant positive relationship of the residential segment indicator with the probability of a false

positive judgment (relative to a correct judgment) was found for interviewers without college education, with no prior NSFG experience, without another job, and working in NSR PSUs only. This positive effect was significantly reduced (suggesting less of an impact of this respondent-level predictor on the probability of a false positive) for interviewers with prior NSFG experience, interviewers with collegiate education, and interviewers working in both SR and NSR PSUs. For interviewers with other jobs or working in both Super 8 PSUs and NSR PSUs, this effect was actually found to become significantly more positive, suggesting that these interviewers tended to struggle when making this judgment in residential segments.

Considering the probability of a false negative judgment relative to a correct judgment, the relationship of the residential segment indicator with the false negative logit was not significant for interviewers with average age, interviewers with an average number of kids, and interviewers not working in all three PSUs. This relationship changed significantly in a *negative* direction for older interviewers, indicating that older age reduced the possible impact of this predictor on the probability of a false negative. Having more kids resulted in a significant *positive* change in this relationship, meaning that interviewers with more kids tended to be more likely to make false negative judgments in residential segments. Finally, this relationship changed significantly in a negative direction for interviewers working in all three PSUs, suggesting that this experience helped with reducing the probability of a false negative when encountering a residential segment.

Next considering interviewer-level moderators of the relationship of evidence of non-English speakers in a segment with judgment accuracy, black interviewers were found to have a significantly decreased relationship of non-English speaking with the probability of a false positive (relative to a correct judgment), relative to interviewers of other races (for whom the relationship of non-English speaking with accuracy was not significant). This suggests that black interviewers working in multi-cultural segments had an easier time judging the presence of children. In addition, interviewers working in SR and NSR PSUs were found to have significantly *increased* odds (relative to those working in other types of PSUs) of making a false positive judgment relative to a correct judgment when working in segments with evidence of non-English speakers. Reasons for this finding are not entirely clear, but working in a mix of SR and NSR PSUs may introduce interviewers to a variety of multi-cultural communities where consistently noticing the absence of children is more difficult. Evidence of non-English speaking was not found to significantly impact the probability of a false negative relative to a correct judgment for any subgroups of interviewers.

Regarding interviewer-level moderators of the effect of reported safety concerns in the segment on judgment accuracy, a non-significant negative effect of safety concerns in the false positive logit for interviewers with mean age and interviewers not working in SR PSUs only was found to significantly change in a positive direction as interviewer age increased. This finding suggests that safety concerns led to more false positive judgments for older interviewers. In addition, the relationship of safety concerns with the false positive logit was found to be become significantly more negative for interviewers

working in SR PSUs only, suggesting that these interviewers had reduced odds of a false positive relative to a correct response when reporting safety concerns. Uncovering mechanisms for this interaction would require discussions with interviewers working in these PSUs. Considering the moderators of the effect of safety concerns in the false negative logit, a marginal negative effect of safety concerns for white and other race interviewers was found to change significantly in a positive direction for black interviewers, with the net effect for black interviewers being close to zero (meaning that safety concerns had little impact on the probability of a false negative for black interviewers). In addition, the effect of safety concerns changed in a significant positive direction for interviewers working in SR and NSR PSUs, suggesting that segments with safety concerns increased the probability of a false negative for interviewers working in both SR and NSR PSUs.

Several interviewer-level features were found to moderate the relationship of observed physical impediments to accessing a housing unit with the probability of a false positive relative to a correct judgment. To begin with, a significant ($p = 0.012$) negative relationship of physical impediments with the odds of a false positive was found for married interviewers who were white or black, had average experience, and were working in NSR PSUs only. This suggests that the odds of a false positive were reduced when physical impediments were encountered for this specific group of interviewers. More experience, being of other ethnicity, and never having been married were interviewer-level features found to significantly change this relationship in a positive direction, suggesting that interviewers with these features had a more difficult time

making this judgment when encountering physical impediments. Interestingly, interviewers working in all three PSU types, SR PSUs only, or Super 8 and SR PSUs (relative to NSR PSUs only) had strongly significant positive changes in this relationship, again suggesting that physical impediments tended to result in more errors in this observation for interviewers working in these PSU types. This finding is illustrated in Figure A.1 below (where the first and second clusters of bars represent interviewers working in NSR PSUs only and Super 8 PSUs only, respectively):

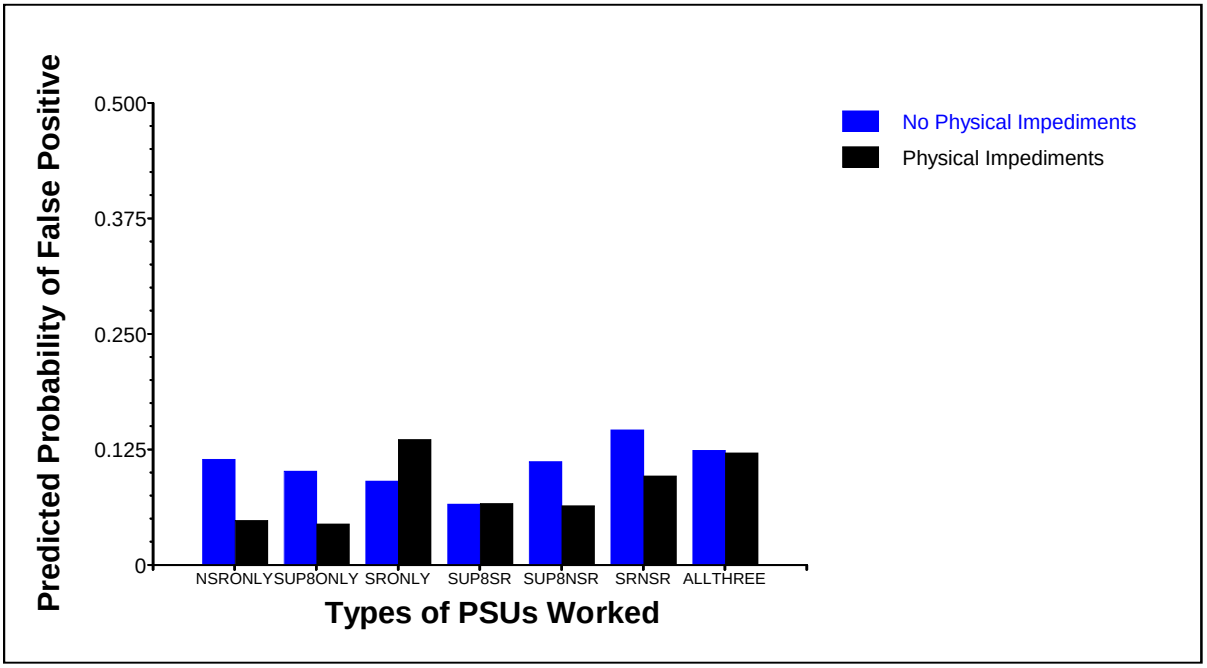


Figure A.1: Predicted probabilities of false positives for judgments on presence of young children for combinations of physical impediments to housing unit access and variety of PSUs worked, based on the exploratory modeling.

Why interviewers working in NSR and Super 8 PSUs only had reduced probabilities of making false positive judgments on the presence of young children when encountering physical impediments relative to interviewers working in other PSU types (especially all three PSU types, or the far-right cluster of bars in Figure A.1, and only SR PSUs, or the

third cluster of bars in Figure A.1) is not clear, but discussions with interviewers about this issue is certainly warranted, given that there is clear variance across types of PSUs worked in terms of the impact of physical impediments on the error rates for this observation. Interestingly, very similar interactions and patterns were found in terms of predicted probabilities of false *negative* judgments relative to correct judgments, suggesting that the relationships of physical impediments with the probability of making an error in general were certainly moderated by the types of PSUs worked by an interviewer.

The relationships of interviewer NSFG (Cycle 7) experience (in days since beginning data collection) at the time of making a judgment with both logits also appeared to be significantly moderated by selected interviewer-level features. First considering the false positive logit, NSFG experience at the time of the interview had a negative but non-significant ($p = 0.10$) relationship with the odds of making a false positive judgment for married white or black interviewers with average prior interviewing experience (NSFG or otherwise), average age, an average number of kids, no college education, and working in NSR PSUs only. More prior interviewing experience, older age, being of other ethnicity, having more kids, working in both Super 8 and SR PSUs, and college education all significantly *increased* the relationship of NSFG experience at the time of the interview with the odds of making a false positive judgment (i.e., the estimated coefficient became more positive), suggesting that these features shifted the coefficient in the direction of more Cycle 7 NSFG experience increasing the probability of a false positive judgment. These findings suggest that those with more experience and more education were at

increased risk of making a false positive judgment farther out in data collection, possibly suggesting laziness on the part of these more experienced and more educated interviewers as data collection proceeds (consistent with findings from McCulloch et al., 2010). In contrast, never having been married, working in Super 8 PSUs only, working in SR PSUs only, and working in Super 8 and NSR PSUs together all significantly *decreased* the relationship of NSFG experience with the false positive logit, suggesting that interviewers with these features tended to improve their observation accuracy as the data collection proceeded. Figure A.2 below illustrates some of these interactions, considering days of interviewer experience at the time of the judgment (a respondent-level feature), college education (an interviewer-level feature), and variety of PSUs worked (an interviewer-level feature).

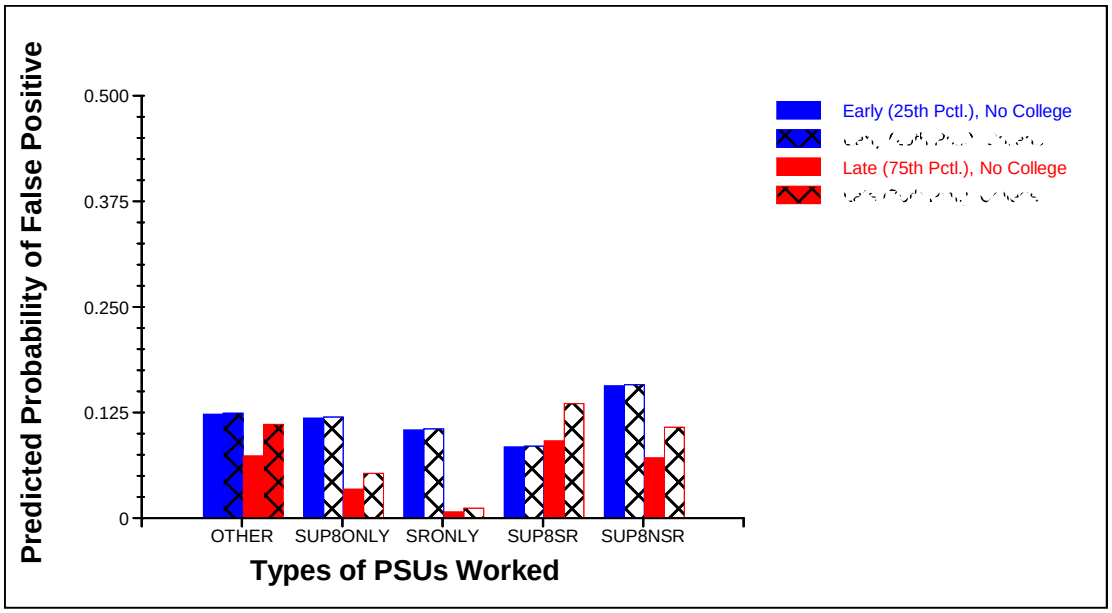


Figure A.2: Predicted probabilities of false positives for judgments on presence of young children for combinations of days of experience since onset of Cycle 7 data collection (Early / Late), interviewer education (College / No College) and variety of PSUs worked, based on the exploratory modeling.

Figure A.2 clearly shows how interviewers working in Super 8 PSUs only, SR PSUs only, and combinations of Super 8 and NSR PSUs have significantly reduced odds of making a false positive judgment as data collection proceeds, relative to interviewers working in other types of PSUs. Figure A.2 also shows the detrimental impact of college education on the change in accuracy as data collection proceeds, where the decrease in the probability of a false positive as data collection proceeds is not as substantial as observed for those interviewers without a college education. Importantly, these results all pertained to the logit for a false positive judgment (relative to a correct judgment).

When considering the logit for a false negative judgment (again relative to a correct judgment), a marginally significant ($p = 0.06$) positive relationship of days of experience was found with the probability of a false negative, specifically for interviewers with average interviewing experience and interviewers not working in both Super 8 and NSR PSUs or all three types of PSUs. Relative to these interviewers, the relationship of days of experience with the false negative logit was significantly decreased (i.e., changed in a negative direction) for interviewers with more interviewing experience, interviewers working in Super 8 and NSR PSUs, or interviewers working in all three PSUs. This suggests that working in a variety of PSUs or higher levels of previous interviewing experience served to attenuate a potentially positive relationship of days of experience with the probability of making a false negative judgment.

Regarding interviewer-level moderators of the differences between segments with higher proportions of minority persons [domain 2, or segments with >10% black residents;

domain 3, or segments with >10% Hispanic residents; and domain 4, or segments with >10% black and >10% Hispanic residents] and segments with higher proportions of white persons (domain 1), three interviewer-level variables were found to significantly ($p < 0.05$) moderate the differences between domain 2 and domain 1 in the false positive logit function. Among white/other interviewers with no college education and no other jobs, there was a slight (and non-significant) decrease in the odds of a false positive for domain 2 relative to domain 1. Interestingly, for black interviewers and interviewers with college education, this decrease changed in a significant manner and became an *increase* in the odds of a false positive, although the net effect was still small. Interviewers with other jobs were found to have a significant *negative* change in this effect, suggesting that interviewers with other jobs had further reduced odds of a false positive in domain 2 relative to domain 1. No interviewer-level variables were found to moderate differences between domain 2 and domain 1 in terms of the odds of a false negative relative to a correct judgment.

The differences between domain 3 (large Hispanic population) and domain 1 in false positive rates were found to vary significantly ($p < 0.01$) depending on college education (where interviewers with college education had significantly reduced probabilities of false positives in domain 3 relative to domain 1), working in combinations of Super 8 and SR PSUs (interviewers working in these PSUs had significant *reduced* odds of making a false positive judgment relative to a correct judgment in domain 3 relative to domain 1), and working in combinations of SR and NSR PSUs (interviewers working in these PSUs had significantly increased differences between domain 3 and domain 1 in the odds of

making a false positive judgment relative to a correct judgment). Interestingly, interviewers of other ethnicity were found to have differences between domain 3 and domain 1 (in terms of the odds of a false positive relative to a correct judgment) that were significantly ($p < 0.05$) *larger* than the differences for white or black interviewers, suggesting that interviewers of other ethnicity tended to struggle with false positives in largely Hispanic segments. Other interviewers were also found to have significantly increased odds of a false *negative* (relative to a correct judgment) in domain 3 relative to domain 1, indicating that interviewers of other ethnicity tended to struggle with this judgment in general in domains that were largely Hispanic. No other interviewer-level features were found to moderate the differences between domain 3 and domain 1 in the false negative logit function.

Finally, regarding the differences between segments in domain 4 (high black *and* high Hispanic proportions) and segments in domain 1 in error rates, a significant ($p < 0.01$) positive difference in false positive rates for those with no prior NSFG experience and not working in combinations of Super 8 and SR PSUs or SR and NSR PSUs was found to be significantly ($p < 0.01$) reduced for those with prior NSFG experience, and significantly *increased* for those working in both Super 8 and SR PSUs or both SR and NSR PSUs (suggesting that those working in these mixes of PSUs generally had a very difficult time with this judgment when working in segments falling into domain 4, potentially due to the variety of situations observed). Figure A.3 below illustrates these differences in expected false positive probabilities.

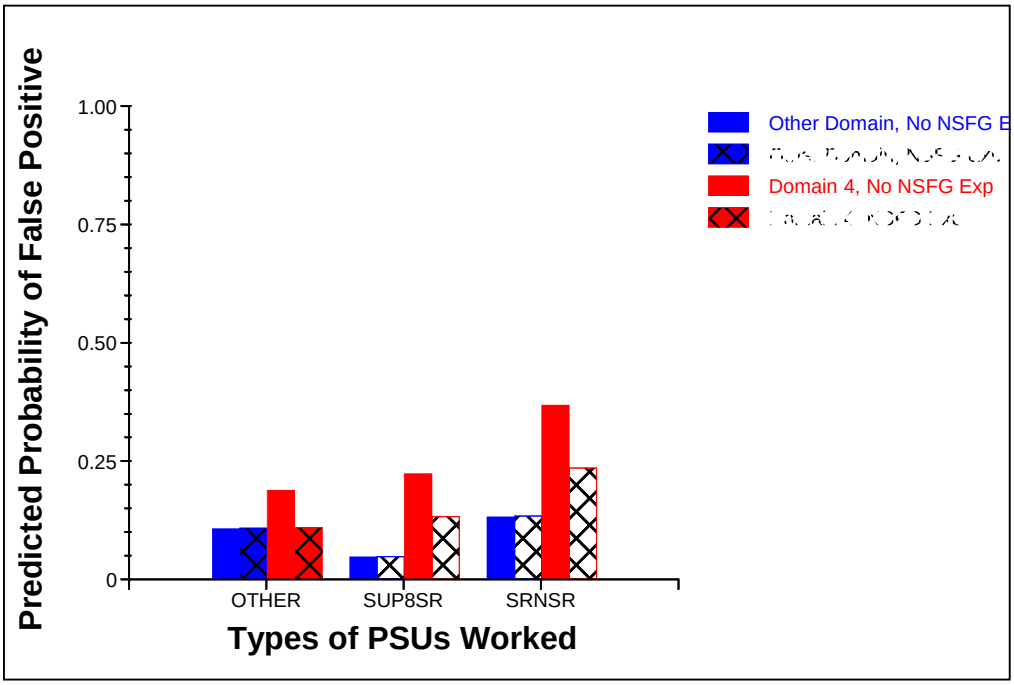


Figure A.3. Predicted probabilities of false positives for judgments on presence of young children for combinations of prior NSFG experience and variety of PSUs worked, based on the exploratory modeling.

Figure A.3 shows the general increase in the probabilities of false positive judgments when working in domain 4 relative to other domains. More importantly, this figure also illustrates how this increase is much larger for those interviewers working in both Super 8 and SR PSUs or SR and NSR PSUs, and how the impact of domain 4 on the probability of a false positive is attenuated for those with prior NSFG experience. These findings have strong practical implications, as effective observational methods in domain 4 are clearly needed for improving observation accuracy. No significant moderators of the differences between domain 4 and domain 1 in terms of false negative rates were found. Finally, due to the sheer number of possible pair-wise comparisons between Census divisions (36), additional interactions between interviewer-level features and the Census

division indicators were not tested, but these models permit that flexibility if hypotheses arise in future investigations.

The vast majority of the significant interactions detailed above remained either marginally significant ($p < 0.10$) or significant ($p < 0.05$) when fitting a model including *all* of the interactions identified in the exploratory analysis (rather than testing them for one respondent-level predictor at a time) in addition to the significant interactions presented in Table 4.5. Collectively, these additional exploratory analyses further suggest that these models can be used to identify combinations of respondent- and interviewer-level features from past data collections that have resulted in increased (or decreased) accuracy in these types of interviewer observations. The results presented in this section could be used by NSFG staff to identify interviewers with specific features who were found to have reduced error rates when faced with certain respondent-level characteristics. NSFG staff could then discuss the observational strategies used by interviewers working in these specific scenarios (e.g., interviewers working in Super 8 PSUs only or NSR PSUs only who had significantly reduced false positive rates for the young children observation when encountering physical impediments to housing unit access, relative to housing units without physical impediments; see Figure A.1), and discern strategies that would be useful for future training sessions.

Additional Exploratory Analyses: Accuracy of Current Sexual Activity Judgments. Consistent with the analyses of the predictors of accuracy in the housing unit observations on young children, several additional exploratory analyses were conducted

to examine potential cross-level interactions of respondent-level predictors with coefficients that were initially found to significantly vary across a *reduced* number of interviewers (due to a lack of consistent within-interviewer variance in the predictors) with interviewer-level features. There was insufficient variance within at least eight interviewers for the following respondent-level predictors also showing evidence of interviewer variance in their coefficients: the intervention indicator (only 42 interviewers with sufficient variance, working in both quarters 1-14 and quarters 15-16), the largely black domain indicator (only 63 interviewers), and the largely Hispanic domain indicator (only 57 interviewers). Of particular interest are interviewer-level moderators of the effects of the intervention on the two logit functions.

No variance was found among interviewers in terms of the effect of the intervention on the probability of a false positive judgment relative to a correct judgment, and as expected, none of the interviewer-level covariates were found to moderate this effect. Based on 42 interviewers, there was evidence of significant variance among interviewers in terms of the effect of the intervention on the probability of a false negative judgment relative to a correct judgment. Unfortunately, none of the available interviewer-level covariates were found to significantly moderate the effects of the intervention in this logit function. Additional research into reasons for the varying effectiveness of this intervention on the probability of making false negative judgments relative to correct judgments is certainly needed, and this may involve in-depth discussions with interviewers found to have unusual predicted values of random effects for this logit.

Exploratory analyses designed to identify interviewer-level predictors of the observed variance in the effects of the largely black segment domain indicator could not be performed, as HLM was unable to invert the resulting variance-covariance matrix of the random effects when all interviewer-level predictors were included in the model. Notably, the variance of the random black domain effects was only weakly significant in the false negative logit ($p = 0.032$) and non-significant in the false positive logit, which may have led to estimation difficulties in the exploratory model. The same problem was observed when attempting to explore interviewer-level predictors of the observed variance in the effects of the largely Hispanic domain indicator on the false negative logit. No additional exploratory analyses were considered, given that few respondent-level predictors had effects on the two logit functions that were found to randomly vary across interviewers.

Appendix B: Cluster Analysis Results from Chapter 4

Figure B.1: Dendrogram showing results of initial hierarchical agglomerative cluster analysis, with evidence of two outliers (Interviewers 36 and 40).

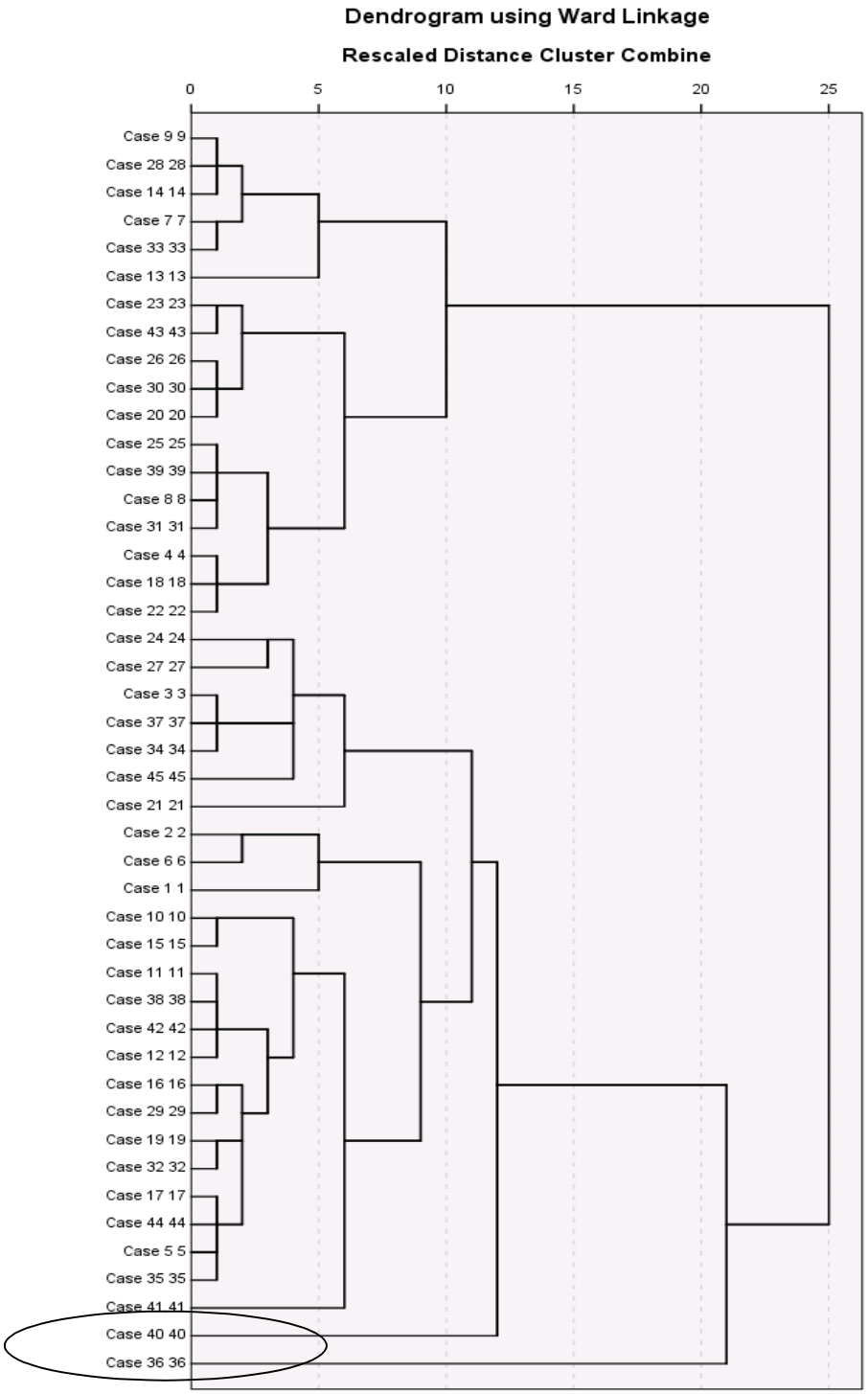


Figure B.2: Dendrogram showing results of second cluster analysis (excluding the two outliers), with evidence of four distinct groups of interviewers (based on rescaled cluster distances greater than 10).

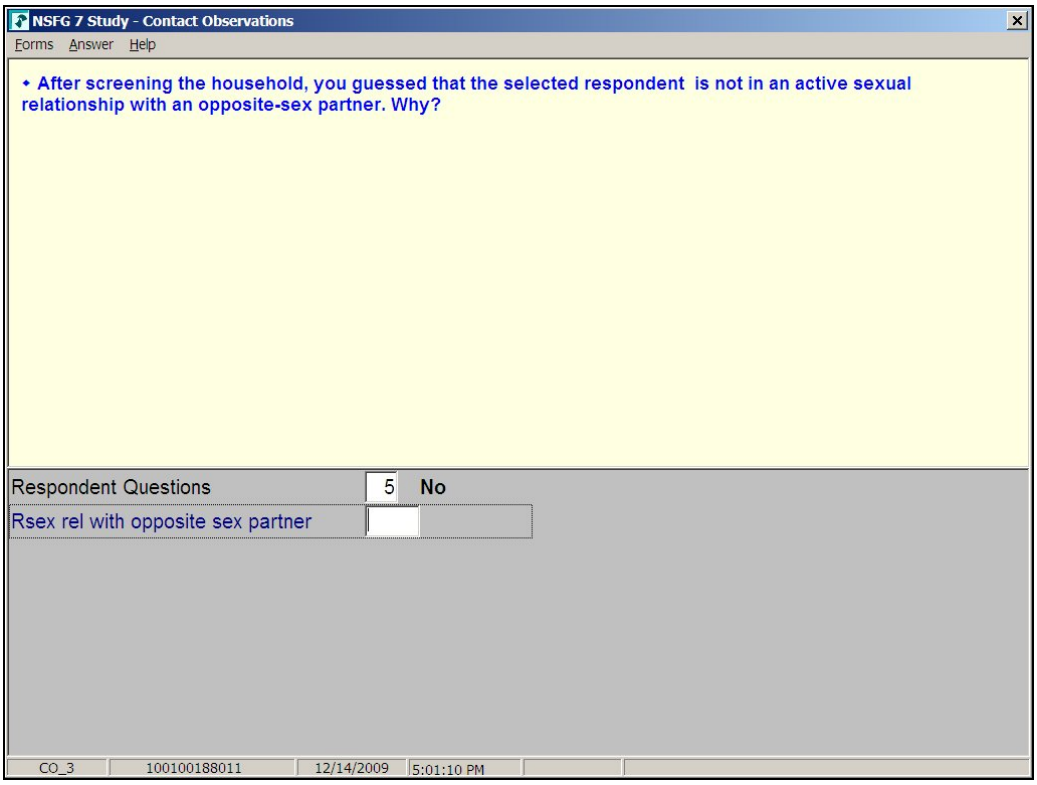


Figure B.3: Screenshot of CAPI application screen where NSFG interviewers could enter open-ended justifications for their sexual activity judgments. The justification was typed into the box labeled "Rsex rel with opposite sex partner," which was converted to a full text box when interviewers started typing.

Appendix C: R Functions Implementing PMM Estimators for Trivariate Normal Incomplete Data

```
##########################
# trivar.analysis function
##########################

# trivar.data is a data frame with n rows and 3 columns,
# and response is a vector with n indicators of response

trivar.analysis <- function(trivar.data,response)
{
      # read in data

      trivar <- trivar.data
      trivar.0 <- trivar[response == 1,]
      trivar.1 <- trivar[response == 0,]

      # sample sizes

      n0 <- length(trivar.0[,1])
      n1 <- length(trivar.1[,1])

      # Draws for Bayesian inference

      pi0draws <- numeric(1000)
      pi1draws <- numeric(1000)
      s22draws <- numeric(1000)
      s11.1.draws <- numeric(1000)
      s11.2.0.draws <- numeric(1000)
      b12.2.draws <- numeric(1000)
      b10.2.draws <- numeric(1000)
      mu.2.0.draws <- numeric(1000)
      mu.1.1.draws <- numeric(1000)
      s33.0.draws <- numeric(1000)
      s33.2.0.draws <- numeric(1000)
      b32.2.draws <- numeric(1000)
      b30.2.draws <- numeric(1000)
      mu2est.draws <- numeric(1000)
      mu3est.draws <- numeric(1000)

      # Regress X1 on X2 for complete cases

      fit12 <- lm(trivar.0[,1] ~ trivar.0[,2])
      b10.2 <- summary(fit12)$coef[1,1]
      b12.2 <- summary(fit12)$coef[2,1]
      b12.2.se <- summary(fit12)$coef[2,2]
      sigma11.2 <- summary(fit12)$sigma^2

      # Regress X3 on X2 for complete cases

      fit32 <- lm(trivar.0[,3] ~ trivar.0[,2])
      b30.2 <- summary(fit32)$coef[1,1]
      b30.2.se <- summary(fit32)$coef[1,2]
      b32.2 <- summary(fit32)$coef[2,1]
      b32.2.se <- summary(fit32)$coef[2,2]
      sigma33.2 <- summary(fit32)$sigma^2

      for (d in 1:1000) # sequence of 12 draws described in the study
```

```
      {
         pi0draws[d] <- rbeta(1, n0 + 0.5, n1 + 0.5)
         pi1draws[d] <- 1 - pi0draws[d]
         s22draws[d] <- n0*var(trivar.0[,2]) / rchisq(1, n0-1)
         while (s11.1.draws[d] <= s11.2.0.draws[d])
         {
            s11.1.draws[d] <- n1*var(trivar.1[,1]) / rchisq(1, n1-1)
            s11.2.0.draws[d] <- n0*sigma11.2 / rchisq(1, n0-2)
         }
         b12.2.draws[d] <- rnorm(1, b12.2, sqrt(s11.2.0.draws[d] /
             (n0*var(trivar.0[,2]))))
         b10.2.draws[d] <- rnorm(1, mean(trivar.0[,1]) –
             b12.2.draws[d]*mean(trivar.0[,2]), sqrt(s11.2.0.draws[d] / n0))
         mu.2.0.draws[d] <- rnorm(1, mean(trivar.0[,2]), sqrt(s22draws[d] /
             n0))
         mu.1.1.draws[d] <- rnorm(1, mean(trivar.1[,1]), sqrt(s11.1.draws[d] /
             n1))
         while (s33.0.draws[d] <= s33.2.0.draws[d])
         {
            s33.0.draws[d] <- n0*var(trivar.0[,3]) / rchisq(1, n0-1)
            s33.2.0.draws[d] <- n0*sigma33.2 / rchisq(1, n0-2)
         }
         b32.2.draws[d] <- rnorm(1, b32.2, sqrt(s33.2.0.draws[d] /
             (n0*var(trivar.0[,3]))))
         b30.2.draws[d] <- rnorm(1, mean(trivar.0[,3]) –
             b32.2.draws[d]*mean(trivar.0[,2]), sqrt(s33.2.0.draws[d] / n0))

         mu2est.draws[d] <- pi1draws[d]*((mu.1.1.draws[d] –
             b10.2.draws[d])/b12.2.draws[d]) + pi0draws[d]*mu.2.0.draws[d]
         mu3est.draws[d] <- b30.2.draws[d] + b32.2.draws[d]*mu2est.draws[d]
      }

      # Compute PMM estimates based on simulated data from posterior
      # distributions

      med.lin.pred <- median(mu2est.draws) # posterior median of X2 means
      med.lin.pred
      pred.prob <- exp(med.lin.pred) / (1 + exp(med.lin.pred))
      med.x3 <- median(mu3est.draws) # posterior median of X3 means

      # For PMM interval coverage and width

      lin.pred.ll <- quantile(mu2est.draws,0.025)
      lin.pred.ul <- quantile(mu2est.draws,0.975)
      pred.prob.ll <- exp(lin.pred.ll) / (1 + exp(lin.pred.ll))
      pred.prob.ul <- exp(lin.pred.ul) / (1 + exp(lin.pred.ul))
      x3.ll <- quantile(mu3est.draws,0.025)
      x3.ul <- quantile(mu3est.draws,0.975)

      cat("PMM estimate of mean of X2 is:"," ",pred.prob,"\n")
      cat("95% PMM CI for mean of X2 is:","
             ",pred.prob.ll,",",pred.prob.ul,"\n")
      cat("95% PMM CI Width is:"," ",pred.prob.ul - pred.prob.ll,"\n")
      cat("\n")
      cat("PMM estimate of mean of X3 is:"," ",med.x3,"\n")
      cat("95% PMM CI for mean of X3 is:"," ",x3.ll,",",x3.ul,"\n")
      cat("95% PMM CI Width is:"," ",x3.ul - x3.ll,"\n")
      cat("\n")
}

# Example run
trivar <- read.csv("J:\\Dissertation\\Paper 3\\nsfg_data.csv")
trivar2 <- trivar[,c(3,8,5)]
```

```
resp <- trivar$response
trivar.analysis(trivar2,resp)



##############################
# trivar.mi.analysis function
##############################

# trivar.data has R rows and 6 columns, response has R indicators of response,
# and m is the number of multiples
# column 1 = X1, column 2 = X2, column 3 = X3, column 4 = strata, column 5 =
# PSU, column 6 = weight)

trivar.mi.analysis <- function(trivar.data, response, m)
{
      # read in data

      trivar <- trivar.data
      trivar.0 <- trivar[response == 1,]
      trivar.1 <- trivar[response == 0,]

      # sample sizes

      n0 <- length(trivar.0[,1])
      n1 <- length(trivar.1[,1])

      # Draws for Bayesian inference

      pi0draws <- numeric(m)
      pi1draws <- numeric(m)
      s22draws <- numeric(m)
      s11.1.draws <- numeric(m) # used in predictive distributions
      s11.2.0.draws <- numeric(m)
      b12.2.draws <- numeric(m)
      b10.2.draws <- numeric(m)
      mu.2.0.draws <- numeric(m)
      mu.1.1.draws <- numeric(m) # used in predictive distributions
      s33.0.draws <- numeric(m)
      s33.2.0.draws <- numeric(m)
      b32.2.draws <- numeric(m)
      b30.2.draws <- numeric(m)
      mu2est.draws <- numeric(m)
      mu3est.draws <- numeric(m)

      # for predictive distributions
      mu.2.1.draws <- numeric(m)
      s12.1.draws <- numeric(m)
      s22.1.draws <- numeric(m)
      mu.3.1.draws <- numeric(m)
      s13.1.draws <- numeric(m)
      s33.1.draws <- numeric(m)
      s32.1.draws <- numeric(m)

      # Regress X1 on X2 for complete cases

      fit12 <- lm(trivar.0[,1] ~ trivar.0[,2])
      b10.2 <- summary(fit12)$coef[1,1]
      b12.2 <- summary(fit12)$coef[2,1]
      b12.2.se <- summary(fit12)$coef[2,2]
      sigma11.2 <- summary(fit12)$sigma^2

      # Regress X3 on X2 for complete cases
```

```
      fit32 <- lm(trivar.0[,3] ~ trivar.0[,2])
      b30.2 <- summary(fit32)$coef[1,1]
      b30.2.se <- summary(fit32)$coef[1,2]
      b32.2 <- summary(fit32)$coef[2,1]
      b32.2.se <- summary(fit32)$coef[2,2]
      sigma33.2 <- summary(fit32)$sigma^2

      s13.2 <- cov(trivar.0[,1],trivar.0[,3]) - cov(trivar.0[,1],trivar.0[,2])
* (cov(trivar.0[,2],trivar.0[,3])/var(trivar.0[,2]))

      for (d in 1:m) # sequence of 12 draws described in the study
      {
            pi0draws[d] <- rbeta(1, n0 + 0.5, n1 + 0.5)
            pi1draws[d] <- 1 - pi0draws[d]
            s22draws[d] <- n0*var(trivar.0[,2]) / rchisq(1, n0-1)
            while (s11.1.draws[d] <= s11.2.0.draws[d])
            {
               s11.1.draws[d] <- n1*var(trivar.1[,1]) / rchisq(1, n1-1)
               s11.2.0.draws[d] <- n0*sigma11.2 / rchisq(1, n0-2)
            }
            b12.2.draws[d] <- rnorm(1, b12.2, sqrt(s11.2.0.draws[d] /
(n0*var(trivar.0[,2]))))
            b10.2.draws[d] <- rnorm(1, mean(trivar.0[,1]) -
b12.2.draws[d]*mean(trivar.0[,2]), sqrt(s11.2.0.draws[d] / n0))
            mu.2.0.draws[d] <- rnorm(1, mean(trivar.0[,2]), sqrt(s22draws[d] /
n0))
            mu.1.1.draws[d] <- rnorm(1, mean(trivar.1[,1]), sqrt(s11.1.draws[d]
/ n1))
            while (s33.0.draws[d] <= s33.2.0.draws[d])
            {
               s33.0.draws[d] <- n0*var(trivar.0[,3]) / rchisq(1, n0-1)
               s33.2.0.draws[d] <- n0*sigma33.2 / rchisq(1, n0-2)
            }
            b32.2.draws[d] <- rnorm(1, b32.2, sqrt(s33.2.0.draws[d] /
(n0*var(trivar.0[,3]))))
            b30.2.draws[d] <- rnorm(1, mean(trivar.0[,3]) -
b32.2.draws[d]*mean(trivar.0[,2]), sqrt(s33.2.0.draws[d] / n0))

            mu2est.draws[d] <- pi1draws[d]*((mu.1.1.draws[d] -
b10.2.draws[d])/b12.2.draws[d]) + pi0draws[d]*mu.2.0.draws[d]
            mu3est.draws[d] <- b30.2.draws[d] + b32.2.draws[d]*mu2est.draws[d]

            # computations for predictive distributions
            mu.2.1.draws[d] <- mean(trivar.0[,2]) + (mu.1.1.draws[d] -
mean(trivar.0[,1])) / b12.2.draws[d]
            s12.1.draws[d] <- cov(trivar.0[,1],trivar.0[,2]) +
(var(trivar.1[,1]) - var(trivar.0[,1])) / b12.2.draws[d]
            s22.1.draws[d] <- var(trivar.0[,2]) + (var(trivar.1[,1]) -
var(trivar.0[,1])) / (b12.2.draws[d]^2)
            mu.3.1.draws[d] <- mean(trivar.0[,3]) + b32.2.draws[d] *
(mu.1.1.draws[d] - mean(trivar.0[,1])) / b12.2.draws[d]
            s13.1.draws[d] <- s13.2 +
b12.2.draws[d]*b32.2.draws[d]*s22.1.draws[d]
            s33.1.draws[d] <- sigma33.2 + (b32.2.draws[d]^2) * s22.1.draws[d]
            s32.1.draws[d] <- b32.2.draws[d] * s22.1.draws[d]
      }

      # create vectors to hold estimates from each MI analysis

      x2meanimpvec <- numeric(m)
      x2varimpvec <- numeric(m)
      x3meanimpvec <- numeric(m)
```

```
      x3varimpvec <- numeric(m)

      # impute missing values based on (5.10) and (5.11) in paper

      require(survey)
      for (b in 1:m)
      {
         # impute missing values based on random draws from the predictive
         # distributions
         trivar.1[,2] <- rnorm(length(trivar.1[,2]), mean = mu.2.1.draws[b] +
(s12.1.draws[b] / s11.1.draws[b]) * (trivar.1[,1] - mu.1.1.draws[b]), sd =
sqrt(s22.1.draws[b] - (s12.1.draws[b]^2/s11.1.draws[b])))

         m1 <- t(c(s32.1.draws[b],s13.1.draws[b]))
         m2 <-
solve(rbind(c(s22.1.draws[b],s12.1.draws[b]),c(s12.1.draws[b],s11.1.draws[b])))
         m3 <- m1 %*% m2
         prod <- numeric(length(trivar.1[,3]))
         for (j in 1:length(trivar.1[,3]))
         {
            x2 <- trivar.1[,2]
            x1 <- trivar.1[,1]
            datavec <- c(x2[j],x1[j])
            diff <- datavec - c(mu.2.1.draws[b],mu.1.1.draws[b])
            prod[j] <- m3 %*% diff
         }

         m4 <- m3 %*% c(s32.1.draws[b],s13.1.draws[b])

         trivar.1[,3] <- rnorm(length(trivar.1[,3]), mean = mu.3.1.draws[b] +
prod, sd = sqrt(s33.1.draws[b] - m4))

         # stack data sets
         trivar.imp <- rbind(trivar.0,trivar.1)
         trivar.imp.data <- data.frame(trivar.imp)

         # set survey features, apply survey mean to complete data (no missing
         # data allowed on design features)
         trivar.imp.data2 <- trivar.imp.data[!is.na(trivar.imp.data[,6]) &
!is.na(trivar.imp.data[,5]) & !is.na(trivar.imp.data[,4]),]
         svyd <- svydesign(strata=~trivar.imp.data2[,4],
id=~trivar.imp.data2[,5], weights=~trivar.imp.data2[,6] ,
data=trivar.imp.data2, nest=T, na.rm=T)

         # save estimate and variance for MI analysis
         x2meanimpvec[b] <- coef(svymean(~trivar.imp.data2[,2],svyd,na.rm=T))
         x3meanimpvec[b] <- coef(svymean(~trivar.imp.data2[,3],svyd,na.rm=T))
         x2varimpvec[b] <- SE(svymean(~trivar.imp.data2[,2],svyd,na.rm=T))^2
         x3varimpvec[b] <- SE(svymean(~trivar.imp.data2[,3],svyd,na.rm=T))^2
      }

      # MI Inference

      x2mean <- mean(x2meanimpvec)
      pred.prob <- exp(x2mean) / (1 + exp(x2mean))
      cat("MI-PMM estimate of mean of X2 is:"," ",pred.prob,"\n")

      # For MI confidence interval coverage and width

      mi.var.x2 <- mean(x2varimpvec) + (1 + 1/m) * var(x2meanimpvec)
      mi.var.x3 <- mean(x3varimpvec) + (1 + 1/m) * var(x3meanimpvec)
      df.x2 <- (m - 1) / ((1 + 1/m) * var(x2meanimpvec) / mi.var.x2)^2
      df.x3 <- (m - 1) / ((1 + 1/m) * var(x3meanimpvec) / mi.var.x3)^2
```

```
      x2mean.ll <- mean(x2meanimpvec) - qt(0.975,df.x2)*sqrt(mi.var.x2)
      pred.prob.ll <- exp(x2mean.ll) / (1 + exp(x2mean.ll))
      x2mean.ul <- mean(x2meanimpvec) + qt(0.975,df.x2)*sqrt(mi.var.x2)
      pred.prob.ul <- exp(x2mean.ul) / (1 + exp(x2mean.ul))
      cat("95% MI-PMM CI for mean of X2 is:"," ", pred.prob.ll, ",",
         pred.prob.ul, "\n")
      cat("95% MI-PMM CI Width is:"," ",pred.prob.ul - pred.prob.ll,"\n")
      cat("\n")

      x3mean <- mean(x3meanimpvec)
      cat("MI-PMM estimate of mean of X3 is:"," ",x3mean,"\n")

      x3mean.ll <- mean(x3meanimpvec) - qt(0.975,df.x3)*sqrt(mi.var.x3)
      x3mean.ul <- mean(x3meanimpvec) + qt(0.975,df.x3)*sqrt(mi.var.x3)
      cat("95% MI-PMM CI for mean of X3 is:"," ",x3mean.ll,",",x3mean.ul,"\n")
      cat("95% MI-PMM CI Width is:"," ",x3mean.ul - x3mean.ll,"\n")
      cat("\n")

}

# Example run
trivar <- read.csv("J:\\Dissertation\\Paper 3\\nsfg_data_new.csv")
trivar2 <- data.frame(cbind(trivar[,5],trivar[,7],trivar[,6],
       trivar[,3],trivar[,2],trivar[,8]))
resp <- trivar$response
trivar.mi.analysis(trivar2,resp,m=20)
```

Appendix D: Additional Simulation Results from Chapter 5

Table D.1 presents a replication of the simulation study (with incomplete data generated from the normal selection model) with $\alpha = -1$ in the response propensity model, which serves to introduce lower expected response rates in the simulated samples.

Table D.1: Selected simulation results under the normal selection model, with $\alpha = -1$ in the response propensity model.

ρ	β	Mean RR	Method	$\hat{\mu}_2$ Rel. Bias (%)	$\hat{\mu}_2$ RMSE	$\hat{\mu}_2$ 95% CI Cover.	$\hat{\mu}_2$ 95% CI Mean Width	$\hat{\mu}_3$ Rel. Bias (%)	$\hat{\mu}_3$ RMSE	$\hat{\mu}_3$ 95% CI Cover.	$\hat{\mu}_3$ 95% CI Mean Width
0.9	2	0.65	PMM	-0.22	0.035	0.944	0.137	-0.01	0.045	0.928	0.162
			MI	11.56	0.120	0.058	0.127	1.74	0.179	0.043	0.180
			GW	11.71	0.130	0.395	0.230	1.72	0.178	0.068	0.182
			CC	43.32	0.434	0.000	0.128	2.18	0.221	0.000	0.148
0.9	1	0.50	PMM	-0.05	0.036	0.955	0.143	-0.01	0.048	0.927	0.180
			MI	10.36	0.109	0.153	0.139	1.58	0.165	0.167	0.211
			GW	9.24	0.102	0.607	0.216	1.33	0.142	0.217	0.191
			CC	41.30	0.415	0.000	0.160	2.07	0.211	0.001	0.172
0.9	0	0.27	PMM	0.05	0.039	0.950	0.159	0.03	0.061	0.933	0.223
			MI	0.12	0.040	0.948	0.172	0.04	0.065	0.946	0.314
			GW	0.08	0.038	0.997	0.240	0.03	0.059	0.962	0.240
			CC	0.15	0.059	0.951	0.240	0.04	0.060	0.958	0.240
0.6	2	0.65	PMM	-0.22	0.053	0.947	0.208	-0.02	0.049	0.930	0.180
			MI	31.23	0.314	0.000	0.134	1.70	0.175	0.024	0.166
			GW	31.51	0.317	0.000	0.138	1.73	0.178	0.011	0.155
			CC	43.25	0.434	0.000	0.127	2.16	0.220	0.000	0.148
0.6	1	0.50	PMM	-0.07	0.060	0.953	0.235	-0.02	0.053	0.936	0.203
			MI	28.19	0.285	0.000	0.168	1.53	0.160	0.159	0.202
			GW	28.15	0.284	0.000	0.171	1.52	0.159	0.092	0.179
			CC	41.30	0.415	0.000	0.160	2.05	0.210	0.002	0.172
0.6	0	0.27	PMM	0.02	0.074	0.968	0.307	0.01	0.067	0.946	0.261
			MI	-0.05	0.053	0.959	0.250	0.01	0.062	0.948	0.291
			GW	0.02	0.050	0.981	0.239	0.01	0.059	0.963	0.240
			CC	0.04	0.060	0.957	0.240	0.01	0.061	0.957	0.240

NOTES: $\rho = \text{corr}(X_1, X_2)$, and defines amount of measurement error in X_1; $\alpha = 0$; β determines dependence of missingness on X_2; PMM = pattern-mixture model estimates based on Bayesian inference approach; MI = multiple imputation estimates after regression prediction and application of Rubin's combining rules; GW = global weighting estimates; CC = complete case estimates; CI = confidence / credible (for PMM) interval

Bibliography

Adams, A.M., Evans, T.G., Mohammed, R., and Farnsworth, J. (1997). Socioeconomic stratification by wealth ranking: is it valid? *World Development*, 25(7), 1165-1172.

Albright, L., Kenny, D.A., and Malloy, T.E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, 55, 387-395.

Alwin, D.F. (2008). *Margins of Error: A Study of Reliability in Survey Measurement*. John Wiley and Sons, Hoboken, New Jersey.

Ambady, N. and Gray, H. (2002). On Being Sad and Mistaken: Mood Effects on the Accuracy of Thin-Slice Judgments. *Journal of Personality and Social Psychology*, 83(4), 947-961.

Ambady, N., Hallahan, M. and Conner, B. (1999). Accuracy of Judgments of Sexual Orientation from Thin Slices of Behavior. *Journal of Personality and Social Psychology*, 77(3), 538-547.

Ambady, N., Hallahan, M. and Rosenthal, R. (1995). On Judging and Being Judged Accurately in Zero-Acquaintance Situations. *Journal of Personality and Social Psychology*, 69(3), 518-529.

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256-274.

Andridge, R.R. and Little, R.J.A. (2009). Extensions of Proxy Pattern-Mixture Analysis for Survey Nonresponse. *In: American Statistical Association Proceedings of the Survey Research Methods Section*: 2468-2482.

Andridge, R.R. and Little, R.J.A. (2011). Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, 27(2), 153-180.

Apoyo Opinión y Mercado. (2004). *Informe gerencial de marketing: niveles socioeconómicos Gran Lima 2004*.

Barnard, J. and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948-955.

Baruch, Y. and Holtom, B.C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, 61(8), 1139-1160.

Baskin, R.M., Zuvekas, S.H., and Ezzati-Rice, T.M. (2011). Proxy Pattern-Mixture Analysis of Missing Health Expenditure Variables in the Medical Expenditure Panel Survey. *Paper presented at the 2011 International Total Survey Error Workshop, Quebec, Canada, June 21, 2011*.

Beaumont, J-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31(2), 227-231.

Berger, G., Hank, L., Rauzi, T., and Simkins, L. (1987). Detection of sexual orientation by heterosexuals and homosexuals. *Journal of Homosexuality*, 13, 83-100.

Bethlehem, J. (2002). Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse* (eds R. Groves, D. Dillman, J. Eltinge and R. Little), pp. 275–287. New York: Wiley.

Biemer, P.P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Biemer, P.P., Chen, P. and Wang, K. (2011). Errors in the Recorded Number of Call Attempts and Their Effect on Nonreponse Adjustments Using Callback Models. *Paper presented at the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland, August 25, 2011*.

Biemer, P.P. and Peytchev, A. (2010). Nonresponse Bias Correction in Telephone Surveys Using Census Geocoding: An Evaluation of Error Properties. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/13/2010.*

Biemer, P.P. and Peytchev, A. (2011). Nonresponse Bias Correction in Telephone Surveys Using Census Geocoding: An Evaluation of Error Properties. *Paper presented at the 2011 International Total Survey Error Workshop, Quebec, Canada, 6/22/2011.*

Biemer, P.P., Wang, K., and Chen, P. (2010). Using call-back data to adjust for nonignorable nonresponse: results of an empirical study. *Paper presented at the 2010 Joint Statistical Meetings, Vancouver, BC, 8/5/2010*.

Biener, L., Garrett, C.A., Gilpin, E.A., Roman, A.M., and Currivan, D.B. (2004). Consequences of Declining Survey Response Rates for Smoking Prevalence Estimates. *American Journal of Preventative Medicine*, 27(3), 254-257.

Blom, A.G. (2009). Nonresponse Bias Adjustments: What Can Process Data Contribute? *Institute for Social and Economic Research Working Paper Series*, No. 2009-21. www.iser.essex.ac.uk.

Blom, A.G., de Leeuw, E.D., and Hox, J.J. (2011). Interviewer Effects on Nonresponse in the European Social Survey. *Journal of Official Statistics*, 27(2), 359-377.

Bolfarine, H. (1991). Finite-Population Prediction under Error-in-Variables Superpopulation Models. *The Canadian Journal of Statistics*, 19(2), 191-207.

Campanelli, P., Sturgis, P., and Purdon, S. (1997). *Can you hear me knocking: An investigation into the impact of interviewers on survey response rates*. London: SCPR.

Casas-Cordero, C. (2010a). Assessing the quality of interviewer observations of neighborhood characteristics. *Paper presented at the 2010 International Total Survey Error Workshop*, Stowe, Vermont, June 14, 2010.

Casas-Cordero, C. (2010b). Testing Neighborhood Mechanisms Influencing Participation in Household Surveys. From the Dissertation *Neighborhood Characteristics and Participation in Household Surveys*, University of Maryland-College Park, 2010.

Casas-Cordero, C. and Kreuter, F. (2008). Assessing interviewer observation of neighborhood characteristics for nonresponse adjustments. *Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (3MC). Berlin, Germany. June 28, 2008.

Couper, M.P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 41-49.

Couper, M.P. and Lyberg, L. (2005). The use of paradata in survey research. *Proceedings of the 55th Session of the International Statistical Institute*.

Cull, W.L., O'Connor, K.G., Sharp, S., and Tang, S.S. (2005). Response rates and response bias for 50 surveys of pediatricians. *Health Services Research*, 40(1), 213-226.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69, 87–98.

Daily, G., Yacey, L.T., Xiao, R., Sreckov, A., Link, M., Bourquin, C., and Shuttles, C. (2008). Transitioning to address-based sampling: Results from Nielsen's TV ratings survey pilot. Paper presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.

D'Arrigo, J., Durrant, G.B. and Steele, F. (2011). Analyzing Interviewer Call Record Data Using a Multilevel Multinomial Modeling Approach to Understand the Process

Leading to Cooperation or Refusal. *Paper presented at the 2011 Joint Statistical Meetings, Miami Beach, FL, 8/2/2010*.

de Leeuw, E., and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Chapter 3 in Groves, R.M. et al., *Survey Nonresponse*. Wiley.

Deming, E. (1944). On Errors in Surveys. *American Sociological Review*, 9, 359-369.

Dillard, D. and Ford, B. (1984). Procedures to Adjust for Nonresponse to the June Enumerative Survey. United States Department of Agriculture, Statistical Research Division, Statistical Reporting Service, SRS Staff Report Number 81.

Drury, T.F., Moy, C.S., and Poe, G.S. (1980). Going beyond interviewer observations of race in the National Health Interview Survey. In: Classification Issues in Measuring the Health Status of Minorities. Hyattsville, MD: National Center for Health Statistics.

Durrant, G.B., D'Arrigo, J. and Steele, F. (2011a). Analyzing Interviewer Call Record Data to Understand the Process Leading to Cooperation or Refusal, submitted; an earlier version of this paper is available as an S3RI Methodology working paper, M11/03.

Durrant, G.B., D'Arrigo, J. and Steele, F. (2011b). Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences, *Journal of the Royal Statistical Society, Series A*, 174, 4, 1-21.

Durrant, G.B., Groves, R.M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, 74(1), 1-36.

Durrant, G.B. and Steele, F. (2009): Multilevel Modelling of Refusal and Noncontact Nonresponse in Household Surveys: Evidence from Six UK Government Surveys, *Journal of the Royal Statistical Society, Series A*, 172, 2, 361-381.

Eckman, S. (2011). Inter-lister Disagreement in Housing Unit Listing. *Paper in Preparation, Chapter of Doctoral Dissertation, Joint Program in Survey Methodology*.

Edwards, S., Brick, J.M., Park, R., and Grant, D. (2011). Validity of Questions to Identify Cell-Only Households. *Paper presented at the 2011 Annual Conference of the American Association for Public Opinion Research*, Phoenix, Arizona, May 14, 2011.

Elliott, M.N., McCaffrey, D., Perlman, J., Marshall, G.N., and Hambarsoomians, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly*, 73(1), 56-73.

Energy Information Administration (1997). Survey Methods of the 1997 Residential Energy Consumption Survey. http://www.eia.doe.gov/emeu/recs/recs97/rx97appa.html.

Everitt, B.S., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis, 5th Edition*. Wiley Series in Probability and Statistics.

Fahimi, M. (2010). Enhancing the Computerized Delivery Sequence File for Survey Sampling Applications. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/14/2010.*

Faraway, J.J. (2005). *Linear Models with R*. Chapman & Hall / CRC Press: Boca Raton, FL.

Feldman, J.J., Hyman, H., and Hart, C.W. (1951). A Field Study of Interviewer Effects on the Quality of Survey Data. *Public Opinion Quarterly*, 15(4), 734-761.

Funder, D.C. (1987). Errors and Mistakes: Evaluating the Accuracy of Social Judgment. *Psychological Bulletin*, 101(1), 75-90.

Funder, D.C. (1995). On the Accuracy of Personality Judgment: A Realistic Approach. *Psychological Review*, 102(4), 652-670.

Fuller, W. (1987). Chapter 1: A Single Explanatory Variable. *Measurement Error Models*. Wiley.

Goldstein, H. (1995). *Multilevel Statistical Models, Second Edition*. Kendall's Library of Statistics 3, Edward Arnold, London.

Gonzalez, J.M. and Kreuter, F. (2010). Evaluating the Impact of Interviewer Observed Auxiliary Information in Nonresponse Adjustments. *Paper presented at the JPSM Brown Bag Seminar, December 7, 2010, University of Maryland – College Park.*

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5), 646-675.

Groves, R.M., Cialdini, R.B., and Couper, M.P. (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly*, 56, 475-495.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society - Series A*, 169(3), 439-457.

Groves, R.M. and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5), 849-879.

Groves, R.M., Mosher, W.D., Lepkowski, J. and Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. National Center for Health Care Statistics. *Vital Health Statistics*, 1(48).

Groves, R.M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Groves, R.M., Wagner, J., and Peytcheva, E. (2007). Use of Interviewer Judgments about Attributes of Selected Respondents in Post-Survey Adjustments for Unit Nonresponse: An Illustration with the National Survey of Family Growth. *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, Salt Lake City, UT.

Hahn, R.A., Truman, B.T., and Barker, N.D. (1996). Identifying ancestry: The reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. *Epidemiology*, 7(1), 75-80.

Heeringa, S.G., West, B.T., and Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman & Hall / CRC Press, Boca Raton, FL.

Hess, I. (2009). *The Practice of Survey Research at the Survey Research Center, 1946-2004*. Survey Research Center, Institute for Social Research, University of Michigan.

Heuer, R.J. (1999). *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, United States of America. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/PsychofIntelNew.pdf

Hill, M.E. (2002). Race of the interviewer and perception of skin color: Evidence from the Multi-city Study of Urban Inequality. *American Sociological Review*, 67(1), 99-108.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition*. New York: John Wiley and Sons.

Hox, J. (1998). Multilevel Modeling: When and Why. In: I. Balderjahn, R. Mathar and M. Schader (Eds.). *Classification, data anaylsis, and data highways*. New York: Springer-Verlag, pp. 147-154.

Hubbard, F. and Lepkowski, J.M. (2009). Experian Database Review. *Internal Memorandum to Statistical Design Group from the Survey Research Center (SRC) Technical Infrastructure Group*, Institute for Social Research, Ann Arbor, MI, 7/16/2009.

Kalton, G., and Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society A*, 149, Part 1, 65-82.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, 64, 125-148.

Kennickell, A.B. (2003). Reordering the Darkness: Application of Effort and Unit Nonresponse in the Survey of Consumer Finances. *Proceedings of the Section on Survey Research Methods, 2003 Joint Statistical Meetings*.

Kennickell, A.B., Mulrow, E., and Scheuren, F. (2011). Paradata or Process Modeling for Inference. Paper presented to the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys. Chicago, IL, June 2011.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, 32(2), 133-142.

Kreuter, F. and Casas-Cordero, C. (2010). Paradata. Section II.4 in *Building on Progress: Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences (Volume 1)*. Edited by the German Data Forum (RatSWD). Budrich UniPress Ltd., Opladen and Farmington Hills, MI.

Kreuter, F., Lemay, M. and Casas-Cordero, C. (2007). Using Proxy Measures of Survey Outcomes in Post-Survey Adjustments: Examples from the European Social Survey (ESS). *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, Salt Lake City, UT.

Kreuter, F. and Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods and Research*, 40(2), 311-322.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys. *Journal of the Royal Statistical Society - Series A*, 173, Part 3, 1-21.

Küchenhoff, H., Mwalili, S.M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, 62, 85-96.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M., and Van Hoewyk, J. (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), June 2010.

Lessler, J. and Kalsbeek, W. (1992). Nonresponse: Dealing with the Problem. Chapter 8 in *Nonsampling Errors in Surveys*. Wiley-Interscience.

Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3), 471-483.

Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics*, 6(3), 287-296.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54, 139-157.

Little, R.J.A. (2008). Weighting and Prediction in Sample Surveys. *Presented in the Celebration of the Diamond Jubilee of the Calcutta Statistical Association Bulletin*.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience, Hoboken, New Jersey.

Little, R.J.A, and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31(2), 161-168.

Little, R.J.A and Wang, Y. (1996). Pattern-Mixture Models for Multivariate Incomplete Data with Covariates. *Biometrics*, 52, 98-111.

Lynn, P. (2003). PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents. *Quality and Quantity*, 37, 239-261.

Maitland, A., Casas-Cordero, C., and Kreuter, F. (2009). An Evaluation of Nonresponse Bias Using Paradata from a Health Survey. *Proceedings of the Section on Government Statistics, Joint Statistical Meetings*, Washington, D.C.

Matsuo, H., Billiet, J. Loosveldt, G. and Malnar, B. (2010). *Response‐based quality assessment of ESS Round 4: Results for 30 countries based on contact files.* Leuven: European Social Survey, University of Leuven.

McCulloch, S.K., Kreuter, F., and Calvano, S. (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, 5/14/2010.*

Merkle, D. and Langer, G. (2008). How Too Little Can Give You a Little Too Much: Determining the Number of Household Phone Lines in RDD Surveys. *Public Opinion Quarterly*, 72(1), 114-124.

Moon, N. (1999). Exit polling: a special case. In *Opinion polls: History, theory and pratice*. Manchester: Manchester University Press. Chapter 7, p. 167.

Moore, C.M. (2001). Inattentional blindness: Perception or memory and what does it matter? *Psyche*, 7(2).

Most, S.B., Scholl, B.J., Clifford, E.R., and Simons, D.J. (2005). What you see is what you get: Sustained Inattentional Blindness and the Capture of Awareness. *Psychological Review*, 112(1), 217-242.

Mosteller, F. (1944). The Reliability of Interviewers' Ratings. Part 2, Chapter 7 in *Gauging Public Opinion*, by Hadley Cantril and Research Associates in the Office of Public Opinion Research, Princeton University, 98-106. Princeton, NJ: Princeton University Press.

Nagelkerke, N. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78 (3), 691–692.

Neisser, U. (1979). The control of information pickup in selective looking. In A.D. Pick (Ed.), *Perception and its development: A tribute to Eleanor J. Gibson* (pp. 201-219). Hillsdale, NJ: Erlbaum.

Olson, K. and Peytchev, A. (2007). Effect of interviewer experience on interviewer pace and interviewer attitudes. *Public Opinion Quarterly*, 71(2), 273-286.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162, Part 3, 437-446.

Passini, F.T. and Norman, W.T. (1966). A universal conception of personality structure? *Journal of Personality and Social Psychology*, 4, 44-49.

Patterson, M.L., Foster, J.L., and Bellmer, C.D. (2001). Another look at accuracy and confidence in social judgments. *Journal of Nonverbal Behavior*, 25(3), 207-219.

Patterson, M.L. and Stockbridge, E. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior*, 22(4), 253-263.

Pickering, K., Thomas, R., and Lynn, P. (2003). Testing the shadow sample approach for the English House Condition survey. *Prepared for the Office of the Deputy Prime Minister by the National Centre for Social Research, London, July 2003*.

Pickery, J. and Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse. *Quality and Quantity*, 36, 427-437.

Punj, G. and Stewart, D.W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, 20(2), 134-148.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3(215), 215-232.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society-A*, 169, 805-827.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85-95.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Rosner, B., Spiegelman, D., and Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4), 734-745.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rule, N.O., Ambady, N., Adams, R.B. and Macrae, C.N. (2008). Accuracy and Awareness in the Perception and Categorization of Male Sexual Orientation. *Journal of Personality and Social Psychology*, 95(5), 1019-1028.

Sakshaug, J.W., Couper, M.P. and Ofstedal, M.B. (2010). Characteristics of Physical Measurement Consent in a Population-Based Survey of Older Adults. *Medical Care*, 48(1), 64-71.

Sarndal, C-E. and Lundstrom, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36(2), 131-144.

SAS Institute Inc. (2011). SAS 9.2 Online Help and Documentation: The GLIMMIX Procedure (Default Output). Cary, NC: SAS Institute Inc., 2008-2011.

Scholes, S., Wardle, H., Sproston, K., Erens, B., Griffiths, M., and Orford, J. (2008). Understanding nonresponse to the British Gambling Prevalence Survey 2007. *Prepared for the Gambling Commission by the National Centre for Social Research, October 2008.* http://www.gamblingcommission.gov.uk/pdf/Understanding non-responses to BGPS 2007 – Oct 2008.pdf.

Schwaninger, A. (2003). Training of Airport Security Screeners. *Airport*, S. 11-13, Git Verlag GmbH & Co. KG, Darmstadt, Germany, www.gitverlag.com/go/airport.

Scioch, P. and Bender, S. (2010). Quality and Quantity: Using Administrative Data for Scientific Purposes in Labor Market Research. *Paper presented at the 2010 Annual*

Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/15/2010.

Shardell, M., Hicks, G.E., Miller, R.R., Langenberg, P., and Magaziner, J. (2010). Pattern-mixture models for analyzing normal outcome data with proxy respondents. *Statistics in Medicine*, 29(14), 1522-1538.

Simons, D.J. and Jensen, M.S. (2009). The effects of individual differences and task difficulty on inattentional blindness. *Psychonomic Bulletin and Review*, 16(2), 398-403.

Sinibaldi, J. (2010). Measurement Error in Objective and Subjective Interviewer Observations. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/14/2010.*

Smith, H.J., Archer, D., and Costanzo, M. (1991). "Just a hunch": accuracy and awareness in person perception. *Journal of Nonverbal Behavior*, 15(1), 3-17.

Smith, T.W. (2011). The Report of the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys. *International Journal of Public Opinion Research*, 23(3), 389-402.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study / validation study designs. *Journal of the American Statistical Association*, 95(449), 51-61.

Stähli, M.E. (2010). Examples and experiences from the Swiss interviewer training on observable data (neighborhood characteristics) for ESS 2010 (R5). Paper presented at the NC Meeting Mannheim, 3/31/2011 – 4/1/2011.

Stefanski, L.A., and Carroll, R.J. (1985). Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, 13(4), 1335-1351.

Steiner, P.M., Cook, T.D., and Shadish, W.R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213-236.

Stokes, L. and Lin, D. (2010). Measurement error in dual frame estimation. *Poster presented at the 2010 International Total Survey Error Workshop*, Stowe, Vermont, June 14, 2010.

Stoop, I., Billiet, J., Koch, A. and Fitzgerald, R. (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Wiley.

Sturgis, P. and Campanelli, P. (1998). The Scope for Reducing Refusals in Household Surveys: An Investigation based on Transcripts of Tape-recorded Doorstep Interactions. *Journal of the Market Research Society*, 40(2).

Su, Y., Gelman, A., Hill, J., and Yajima, M. (2009). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*, 20(1), 1-27.

Tipping, S. and Sinibaldi, J. (2010). Examining the trade off between sampling and targeted non-response error in a targeted non-response follow-up. *Paper presented at the 2010 International Total Survey Error Workshop*, Stowe, Vermont, June 15, 2010.

Tolonen, H., Helakorpi, S., Talala, K., Helasoja, V., Martelin, T., and Prattala, R. (2006). 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *European Journal of Epidemiology*, 21, 409-415.

Trappmann, M., Gundert, S., Wenzig, C., and Gebhardt, D. (2011). PASS: a household panel survey for research on unemployment and poverty (im Er-scheinen). In: *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissen-schaften*, Vol. 131, No. 1.

Tucker, C., Brick, J.M., and Meekins, B. (2007). Household Telephone Service and Usage Patterns in the United States in 2004: Implications for Telephone Samples. *Public Opinion Quarterly*, 71(1), 3-22.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, 74(2), 223-243.

Ward, Joe H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

West, B.T. (2010a). A Practical Technique for Improving the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth. NSFG Paper No. 10-013. November 2010.

West, B.T. (2010b). The Impact of Measurement Error in Auxiliary Variables on Model-Based Estimation of Finite Population Totals: A Simulation Study. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Pages TBD. *Winner of the 2010 Student Paper Competition Sponsored by the Social Statistics, Government Statistics, and Survey Research Methods Sections of the American Statistical Association*.

West, B.T. and Little, R.J.A. (2012). Pattern-Mixture Models for Trivariate Normal Incomplete Survey Data with Fully Observed Auxiliary Variables Subject to Measurement Error. *Paper to be presented at the 2012 Federal Conference on Statistical Methodology, Washington, D.C.*

West, B.T., Welch, K.B., and Galecki, A.T. (2007). *Linear Mixed Models: A Practical Guide using Statistical Software*. Chapman Hall / CRC Press.

Winerman, L. (2005). 'Thin Slices' of Life. *Monitor*, 36(3), 54. http://www.apa.org/monitor/mar05/slices.aspx.

Wolfe, J.M. (1999). Inattentional Amnesia. In V. Coltheart (Ed.), *Fleeting Memories: Cognition of brief visual stimuli* (pp. 71-94). Cambridge, MA: MIT Press.

Wolter, K.M. (2007). *Introduction to Variance Estimation (Second Edition)*. Springer-Verlag, New York.

Wolter, K.M., Montogomery, R., Tao, X., Greby, S., and Kennedy, E. (2011). Accuracy of Geographic Stratification in a Cell-Phone Survey. *Paper presented at the 2011 Annual Conference of the American Association for Public Opinion Research*, Phoenix, Arizona, May 15, 2011.

Yan, T., Datta, R., Wolter, K.M., and Shin, H-C. (2011). Use of Paradata to Assess the Performance of a Multi-Mode ABS Study. *Paper presented at the 2011 Joint Statistical Meetings*, Miami Beach, Florida, July 31, 2011.

Yan, T. and Raghunathan, T. (2007). Using Proxy Measures of the Survey Variables in Post-Survey Adjustments in a Transportation Survey. *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, Salt Lake City, UT.

Yi, G.Y. (2011). Analysis of Imperfect Data: Handling Missing Observations and Measurement Error. *Paper presented to the University of Michigan-Ann Arbor Department of Biostatistics, April 7, 2011.*

Yi, G.Y., Liu, W., and Wu, L. (2011). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*, 67(1), 67-75.

Zhang, D. and Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal / clustered data and other related topics. Chapter 2 in *Random effect and latent variable model selection*, ed. D. B. Dunson. Springer Lecture Notes in Statistics, 192.

Zhang, J. (2011). Differential Effects of Measurement Error in Outcome Prediction. *Poster presented at the 2011 Michigan Student Symposium in Interdisciplinary Statistical Sciences, April 8, 2011, Ann Arbor, MI.*